DeepMind

Representation Learning Without Labels

Irina Higgins @irinavlh
Danilo J. Rezende @danilojrezende
S. M. Ali Eslami @arkitus

ICML 2020

DeepMind

Acknowledgements

Mihaela Rosca, Shakir Mohamed, Alex Graves,
Olivier Henaff, Brian McWilliams, Steven McDonagh,
David Pfau, Jovana Mitrovic, Andrew Zisserman

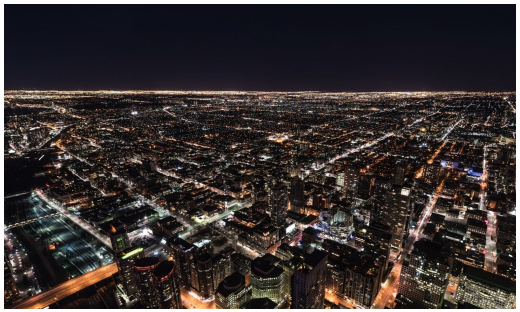
ICML 2020

Agenda for this tutorial

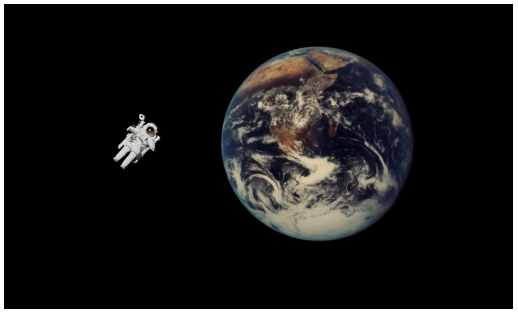
01
Introduction

02
Building Blocks

03
Landscape



04
Frontiers



Scope for this tutorial

What this tutorial is

- An overview of building blocks
- A comparison of methods
- Focus on the image modality

What this tutorial isn't

- A comprehensive list of all relevant techniques
- Representation learning for different modalities (text, audio, video, graphs, etc.)

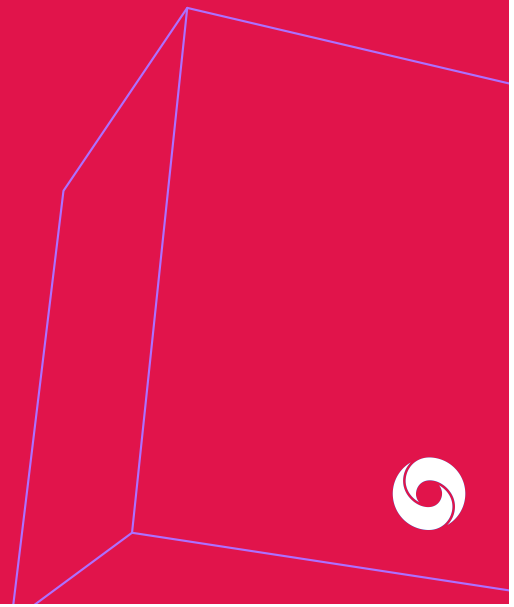
Disclaimers

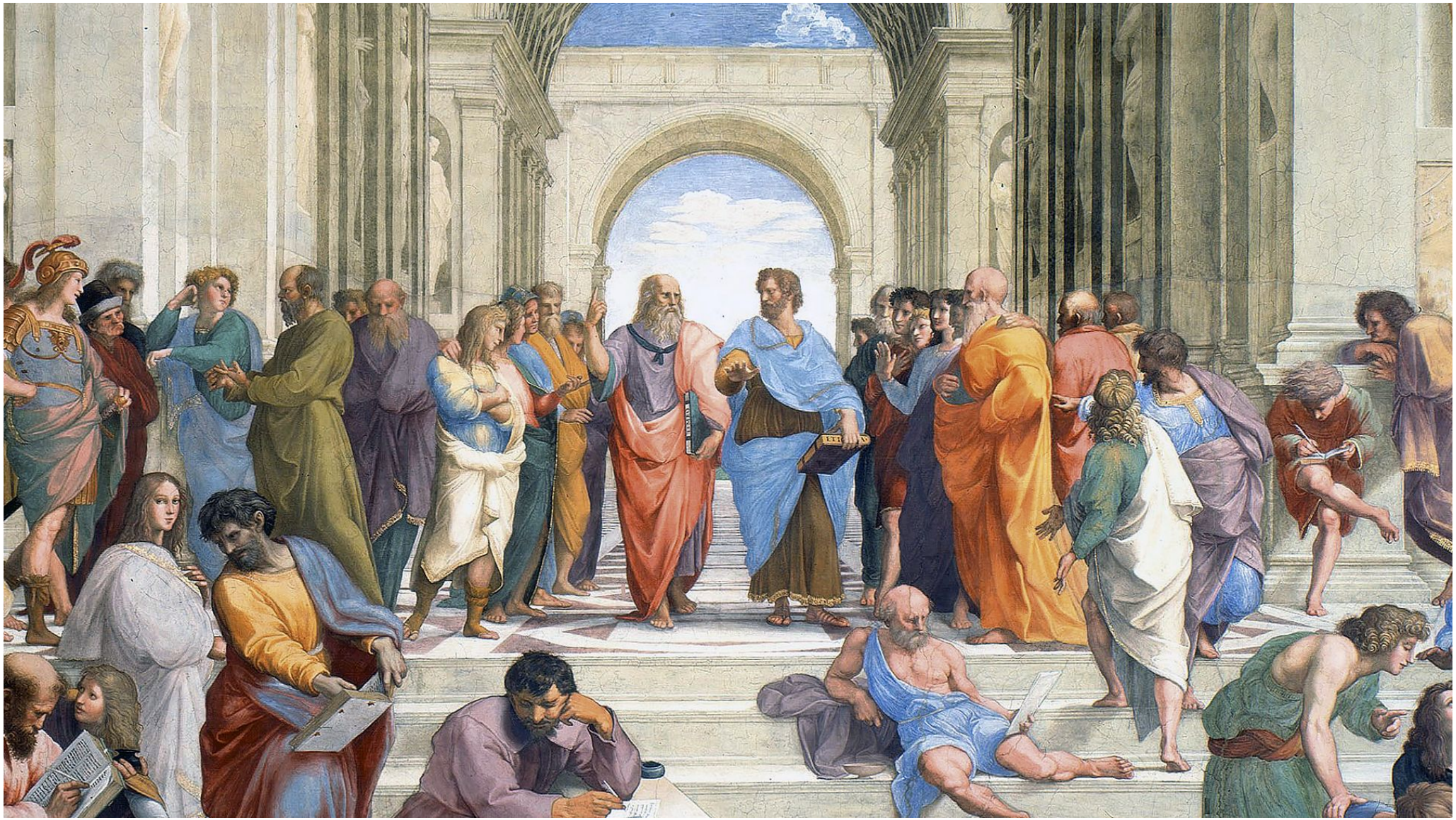
Want to learn more?

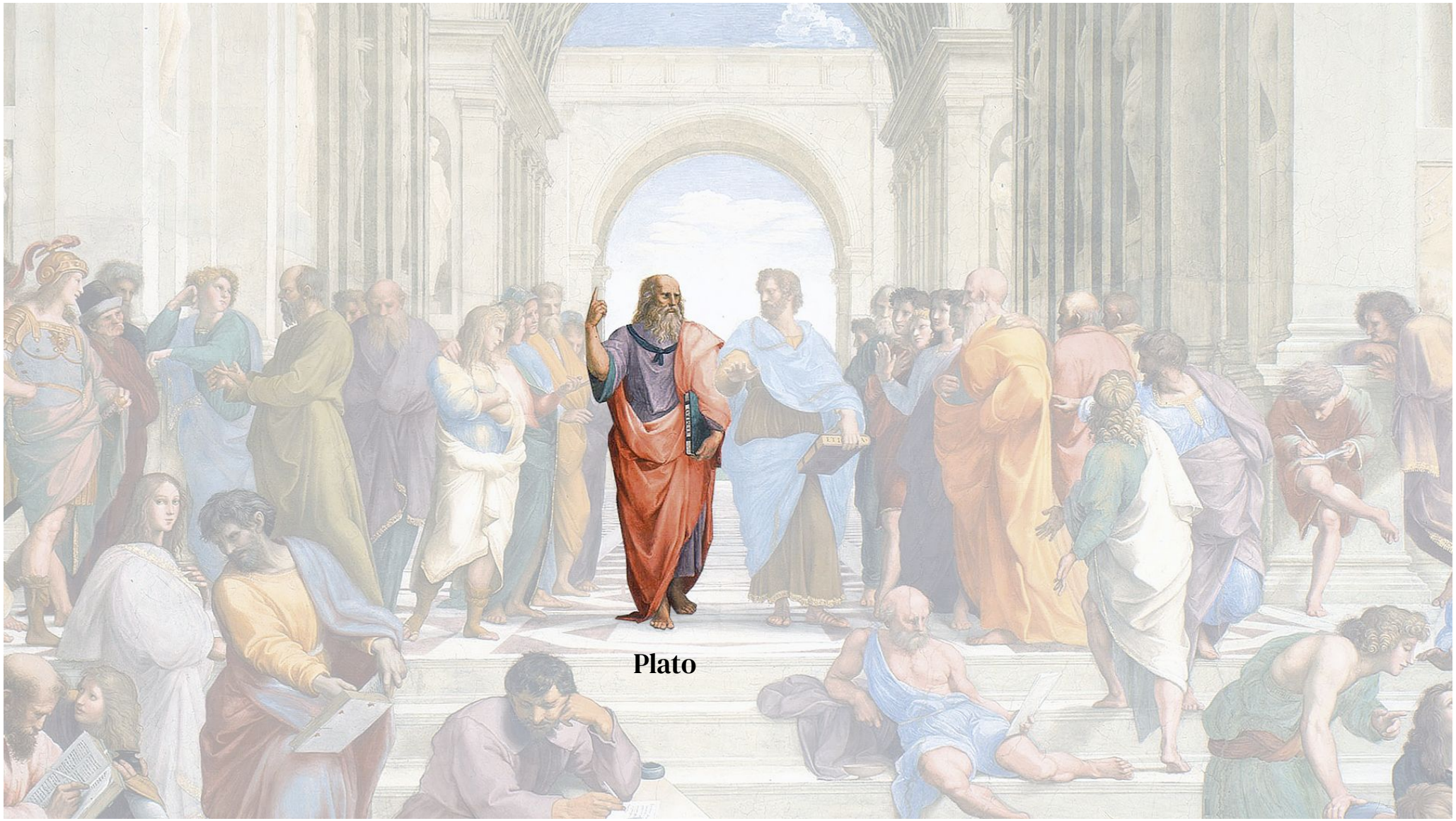
- This is a huge topic with a vast, multi-disciplinary history
- We will inevitably miss important related work
- Each citation is only meant as a representative example; seed for **connectedpapers.com**
- There are many views on the literature, this is one
- Not necessarily chronological
- Email us with pointers or suggestions and we will update the slides

DeepMind

1 Introduction

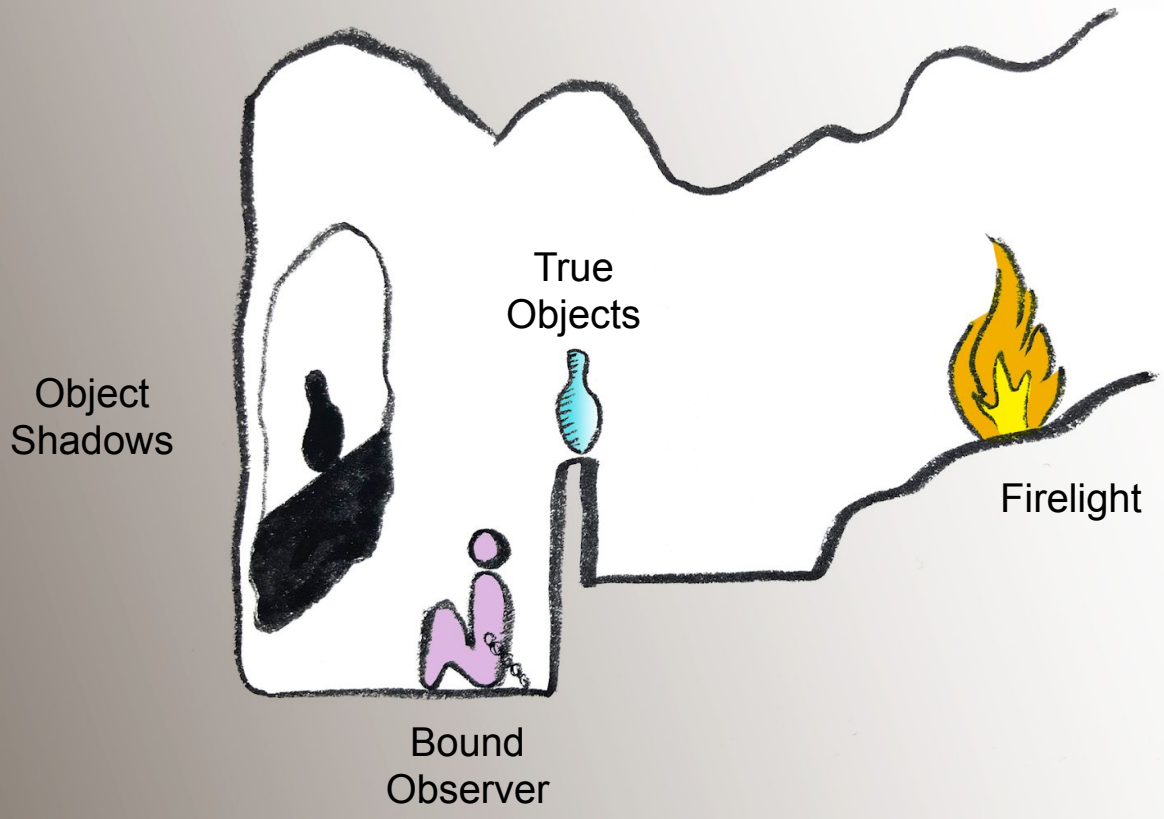




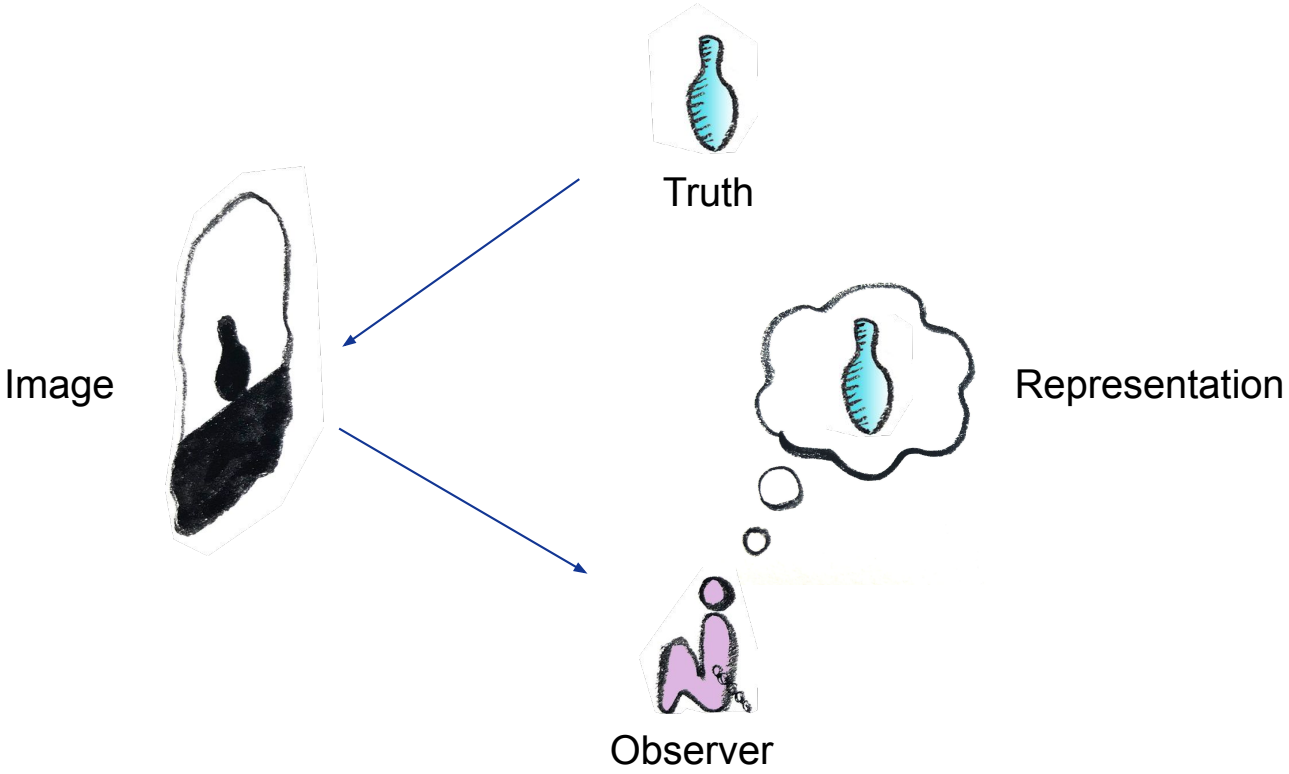


Plato

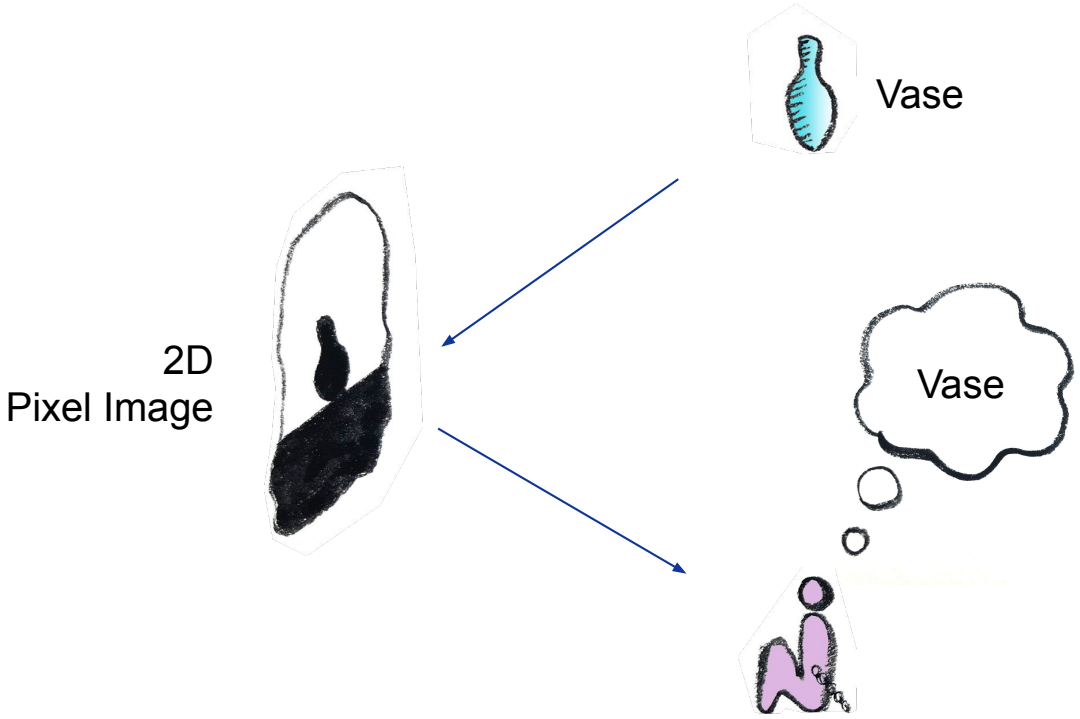
Plato's allegory of the cave



Plato's allegory of the cave

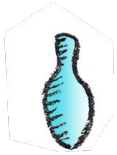


Desired understanding: simplistic view

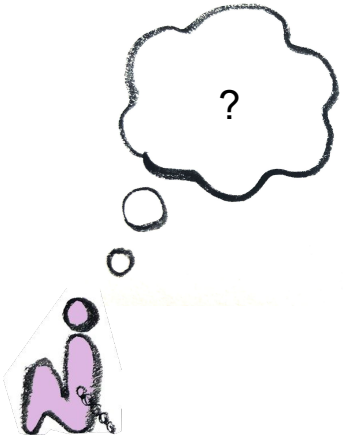


The representation problem

2D
Pixel Image



- **Class:** Vase
- **Shape:** Gourd
- **Colour:** Blue
- **Height:** 15cm tall
- **Weight:** 230g
- **Scratched:** Yes
- **Moving:** No
- And many more attributes...



- Which attributes?
- What formats?
- Partial observability?
- How quickly?
- Measure of success?

> Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered by irrelevant information.

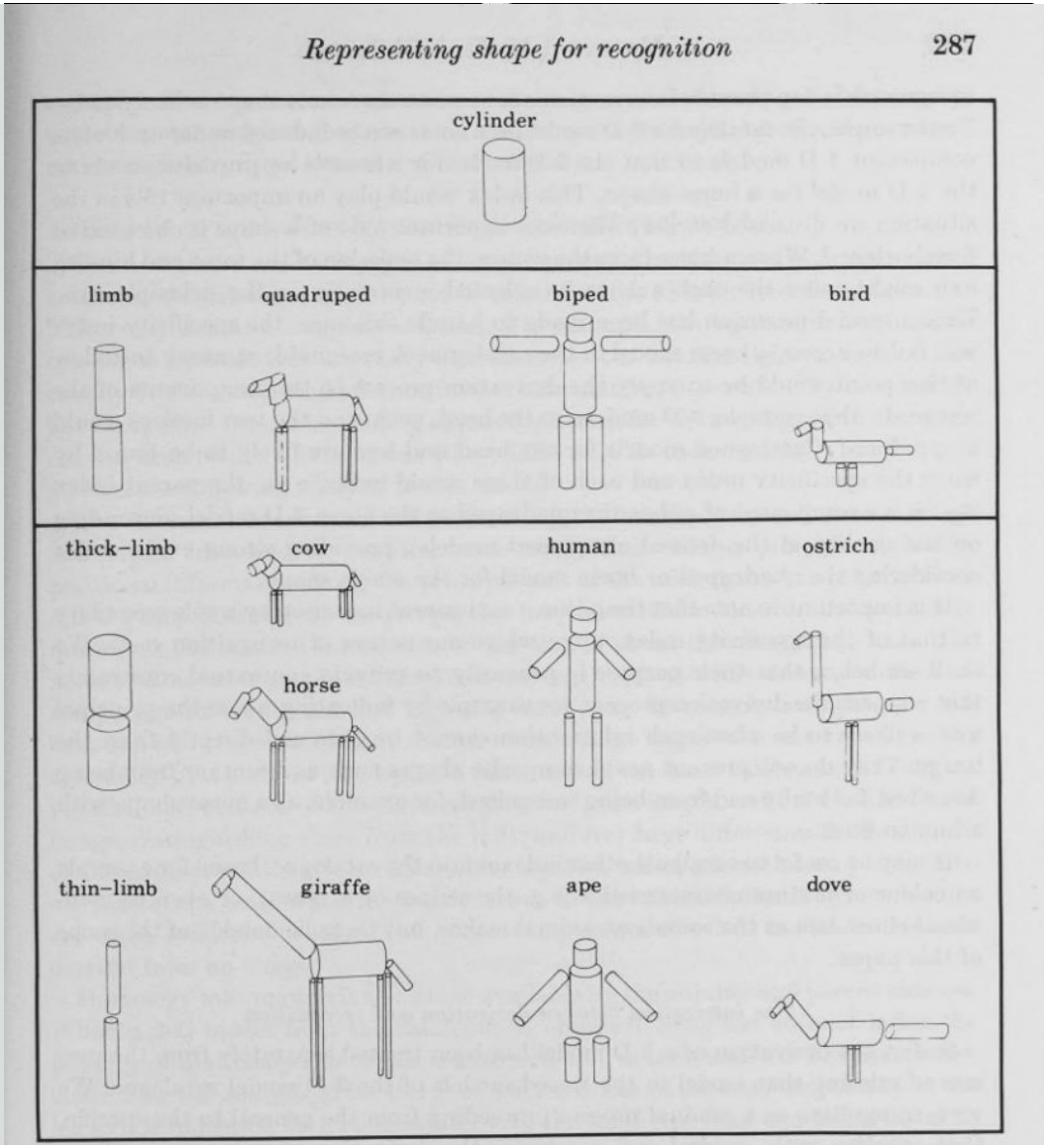
— Marr and Nishihara, 1978

Want to learn more?

Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes, Marr et al, Proc. R. Soc. Lond. (1978)

> **Representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this.**

— Marr and Nishihara, 1978



Want to learn more?

How Can Deep Learning Advance Computational Modeling of Sensory Information Processing? Thompson et al, arxiv (2018)

- Representational **form** is orthogonal to information **content**
- Useful **abstraction** to make different computations and tasks **more efficient**

The supervised solution

1. Decide **which attributes** you care about
2. Decide the **format** for each attribute
3. Create a **large dataset** of (image, label) pairs
4. **Train a neural network** to predict labels

label= ‘Vase’ label= ‘Ball’ label= ‘Dog’

The supervised solution

1. Decide **which attributes** you care about
2. Decide the **format** for each attribute
3. Create a **large dataset** of (image, label) pairs
4. **Train a neural network** to predict labels

Works extremely well on a diverse set of problems

However, it also raises several questions:

1. Who provides ground truth for the labels?
2. What if we don't 'know' the groundtruth ourselves?
3. Which attributes are chosen for labelling?
4. Which attributes are ignored for labelling?
5. What biases do the labels propagate?
6. Do children learn purely from labels?
7. Do animals learn from labels at all?
8. **Can useful representations develop without labels?**

label= 'Vase'

label= 'Ball'

label= 'Dog'

Can useful representations develop without labels?

If the answer is yes:

1. We learn more efficiently when we do gain access to labels

2. We still learn useful things when label collection is impossible

Supervised vs Unsupervised
Labelled vs Unlabelled

Often it is difficult to formally distinguish between **supervised** and **unsupervised** techniques

For example:

- Image captioning (specification of **caption**)
- Machine translation (specification of **pairing**)
- Reinforcement learning (specification of **reward**)
- Generative models (specification of **structure**)
- Language modeling (specification of **dataset**)

Objective is to reduce reliance on labels **that can only be assigned by human brains**, and learn more from **raw measurements of the world**

Still, the term ‘unsupervised’ is actually rather convenient

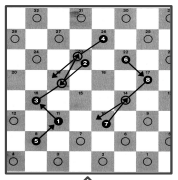
Universal Dictionary of Arts, Sciences and Literature, 1810

History of representation learning

Want to learn more?

Some Studies in Machine Learning Using the Game of Checkers, Samuel, IBM Journal (1959)

- Arthur Samuel coins the term “machine learning”



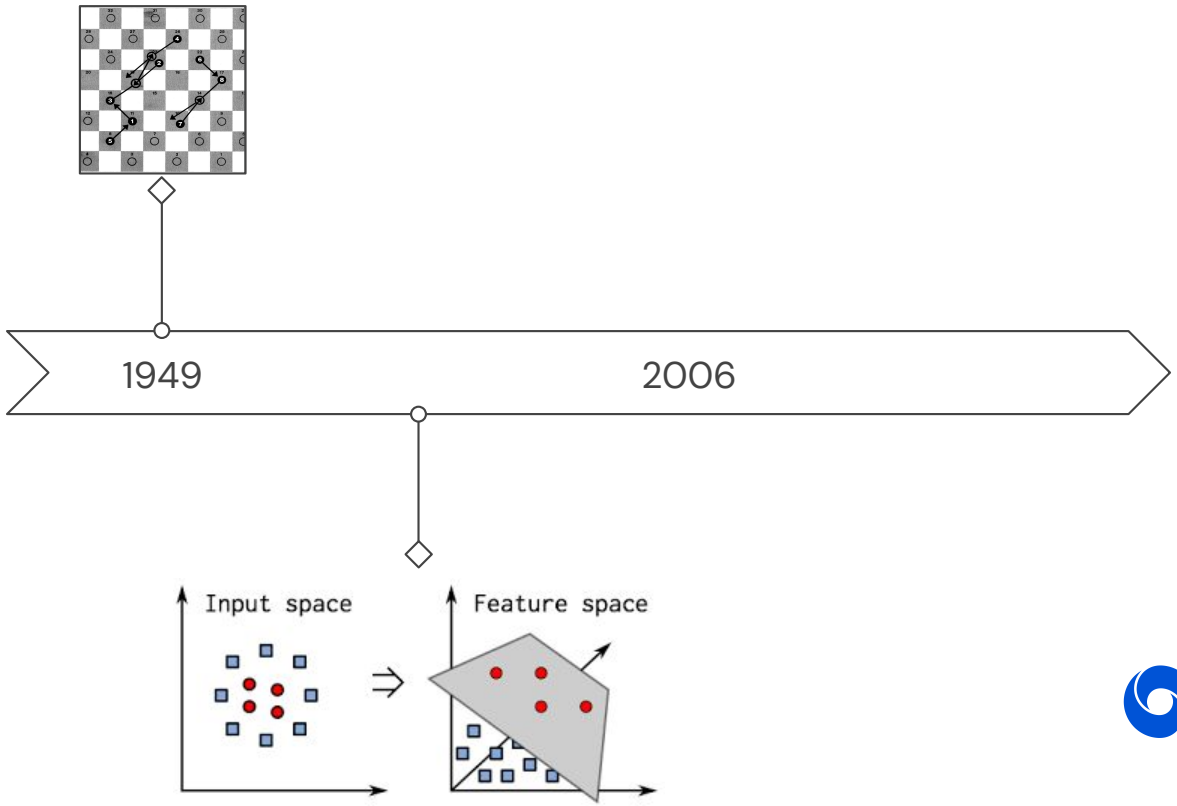
1949

History of representation learning

Want to learn more?

Kernel Methods in Machine Learning, Hofmann et al, The Annals of Statistics (2008)

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods

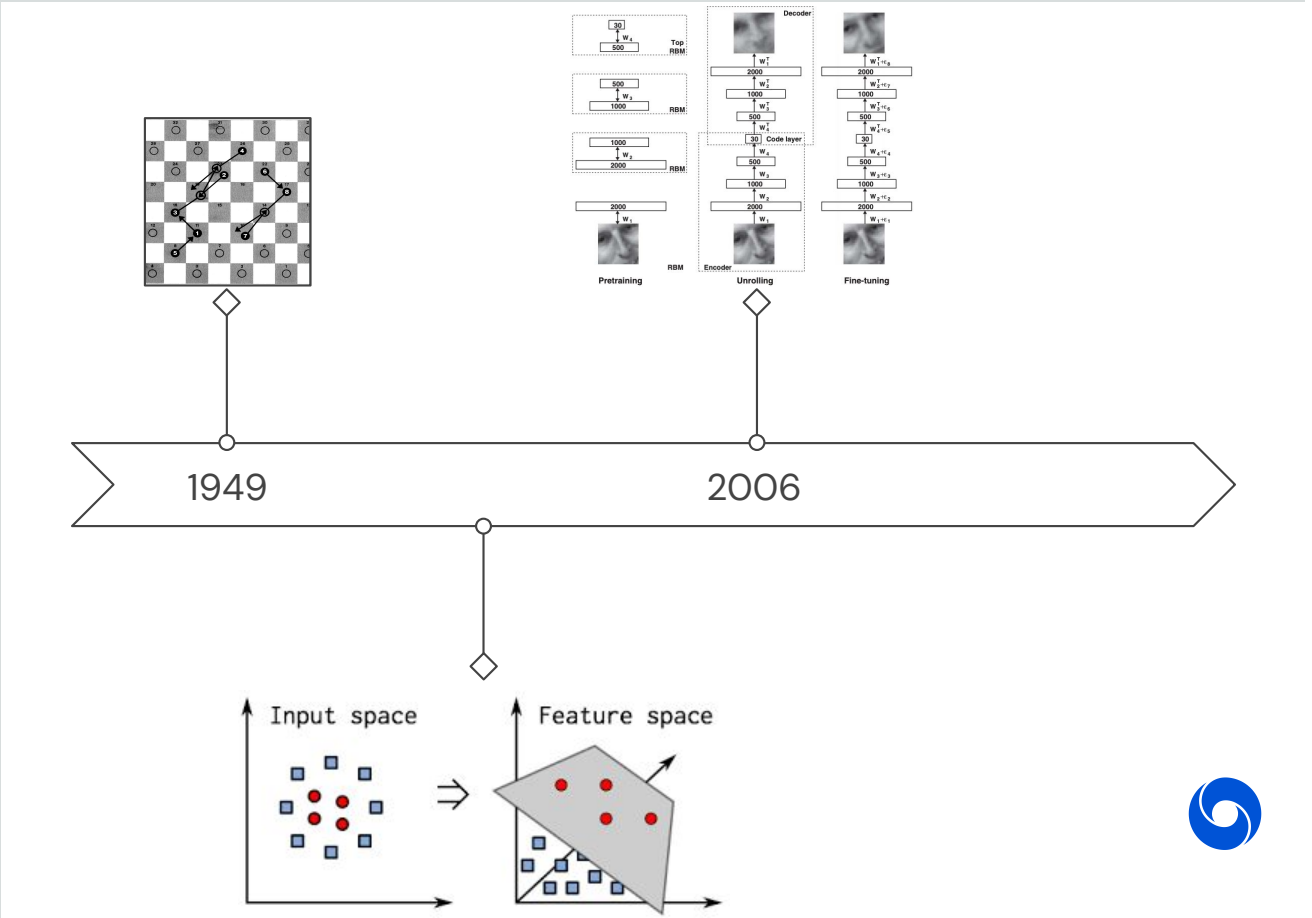


History of representation learning

Want to learn more?

Reducing the Dimensionality of Data with Neural Networks, Hinton and Salakhutdinov, Science (2006)

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods
- Restricted Boltzmann Machines used for initialising deep classifiers

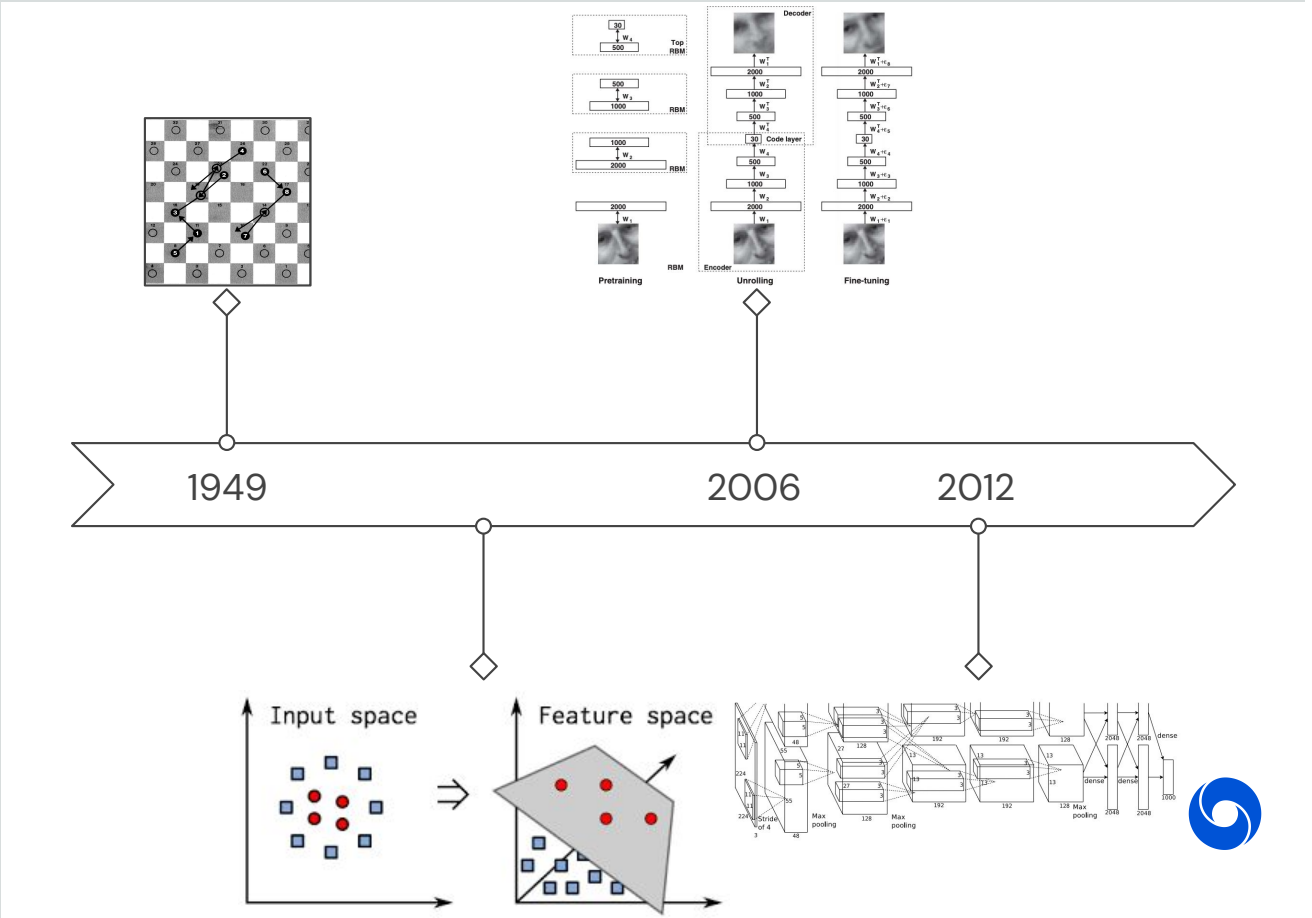


History of representation learning

Want to learn more?

ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky et al, NeurIPS (2012)

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods
- Restricted Boltzmann Machines used for initialising deep classifiers
- AlexNet wins ImageNet challenge by a large margin with no unsupervised pre-training



Turing Award winners at AAAI 2020

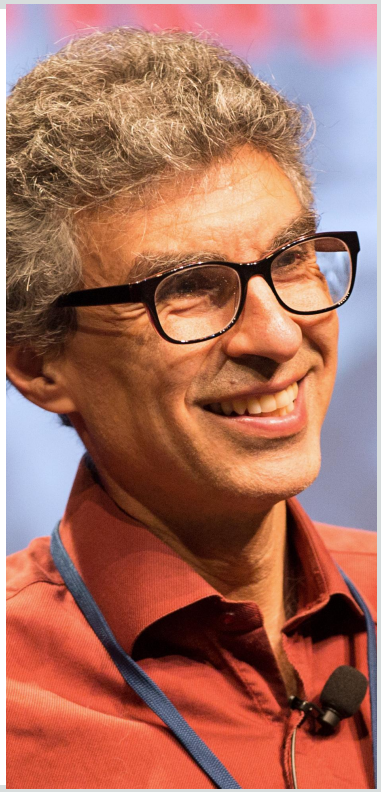
> I always knew unsupervised learning was the right thing to do
>
> — Geoff Hinton

> Basically it's the idea of learning to represent the world before learning a task — and this is what babies do
>
> — Yann LeCun

> And so if we can build models of the world where we have the right abstractions, where we can pin down those changes to just one or a few variables, then we will be able to adapt to those changes because we don't need as much data, as much observation in order to figure out what has changed.
>
> — Yoshua Bengio

Jérémy Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0

Eviatar Bach / CC BY-SA



Jérémy Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0

Intelligence without representation

> “There is no clean division between perception (abstraction) and reasoning in the real world. The brittleness of current AI systems attests to this fact.
>
> — Rodney Brooks

Want to learn more?

Intelligence without representation, Rodney A. Brooks, Artificial Intelligence (1991)

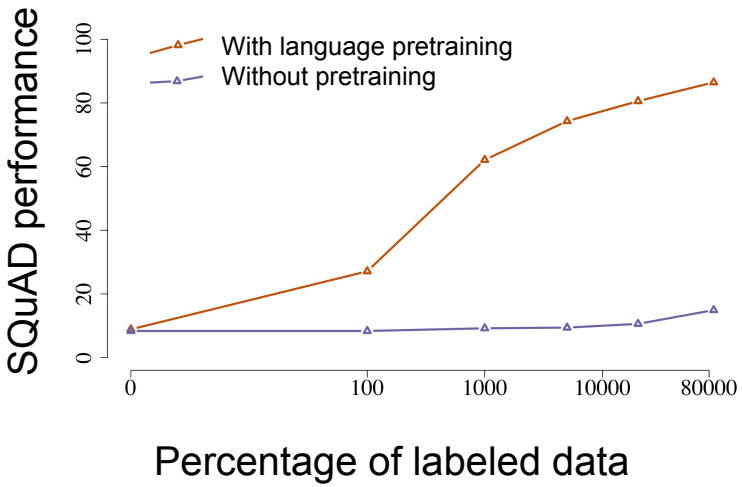
Science History Institute / CC BY-SA

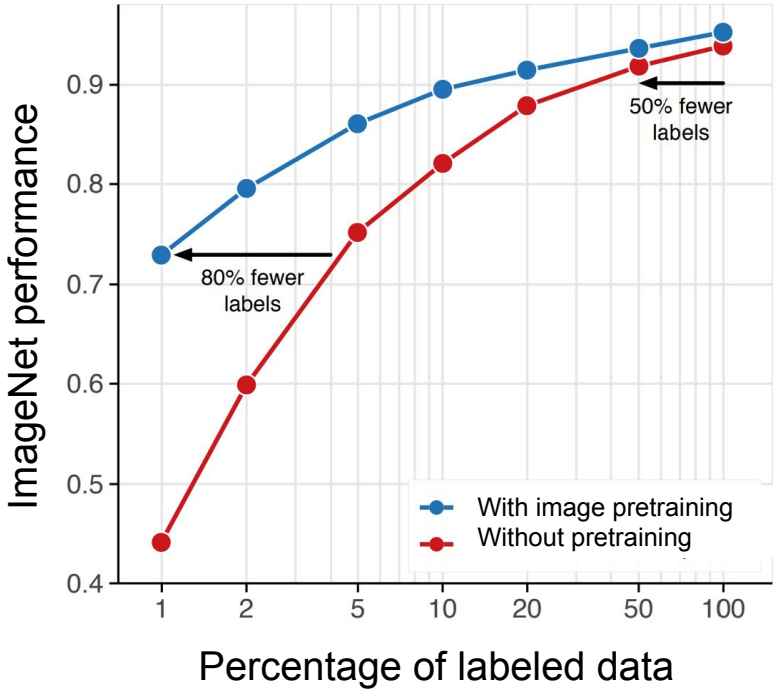
Impressive recent progress

Want to learn more?

Learning and Evaluating General Linguistic Intelligence, Yogatama et al (2019)

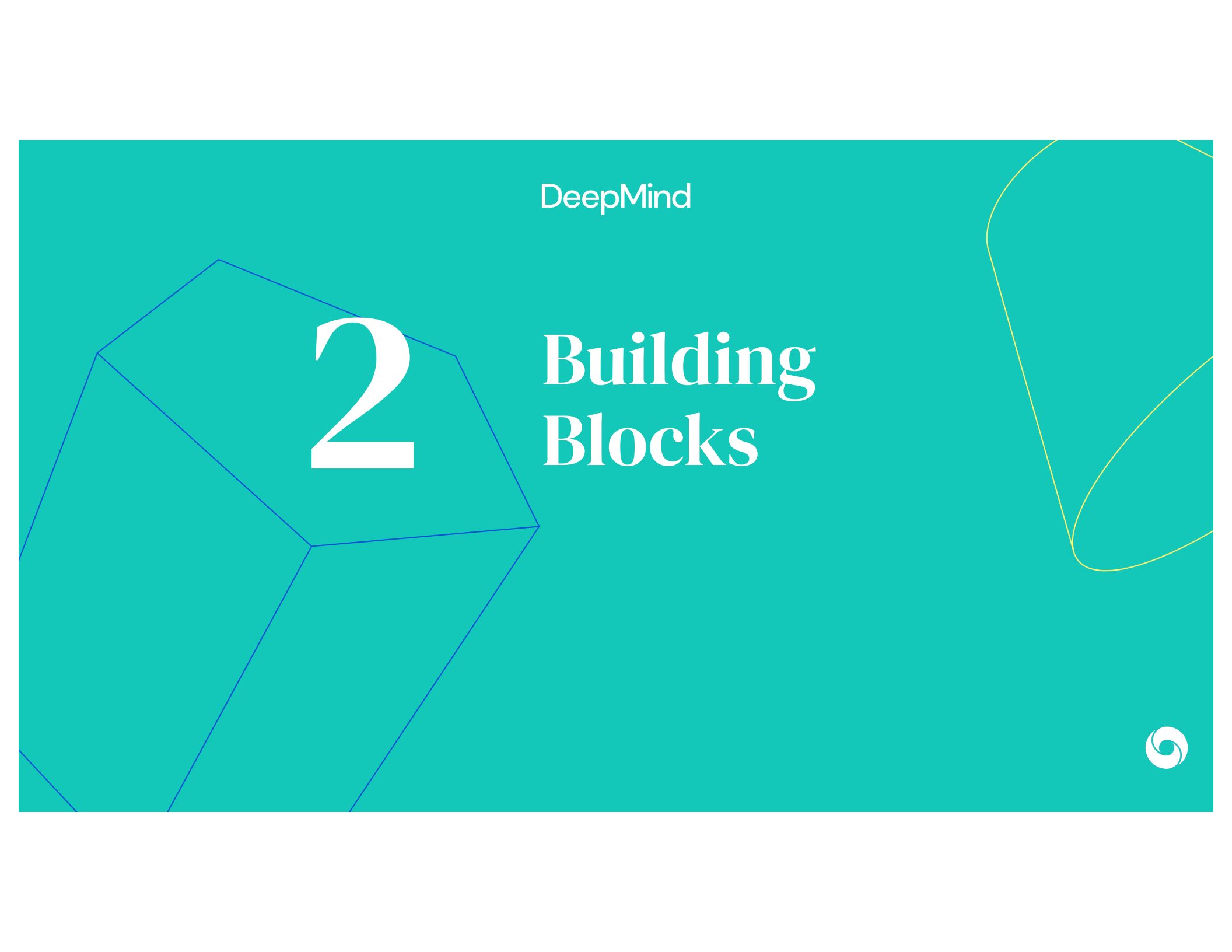
Data-Efficient Image Recognition with Contrastive Predictive Coding, Olivier J. Hénaff et al, ICML (2020)





DeepMind

2 Building Blocks



The representation problem

Realm of
Images

Representation
(Understanding)

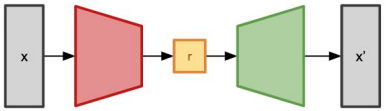
Generation
(Imagination)

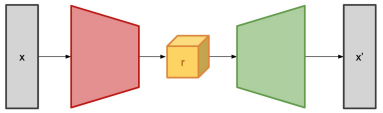
Realm of
Representations

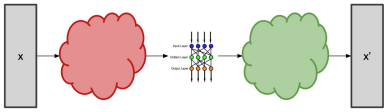
Model zoo

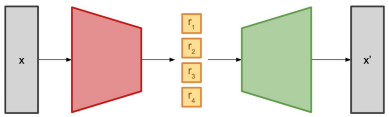
Curtis's Botanical Magazine, 1831

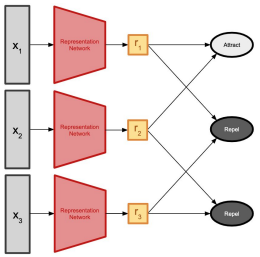
Model zoo

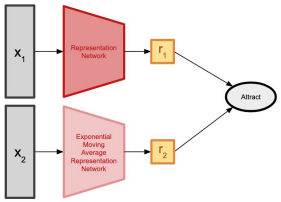

x
r
x'

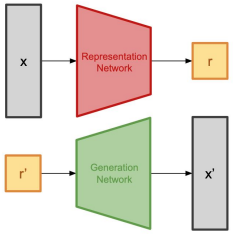

x
r
x'

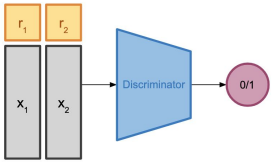

x
x'

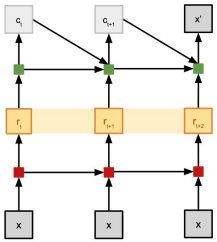

x
x'

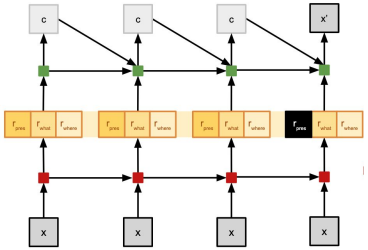

Representation Network
Attract
Repel

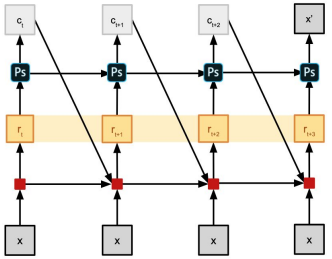

Representation Network
Exponential Moving Average Representation Network
Attract


x
Representation Network
r
r'
Generation Network
x'

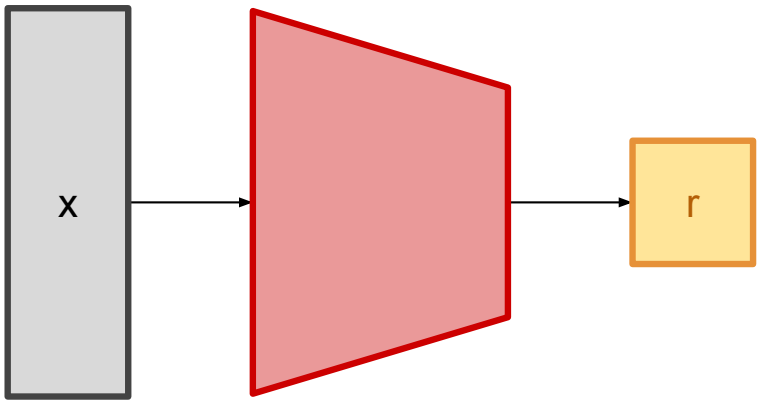

Discriminator
0/1







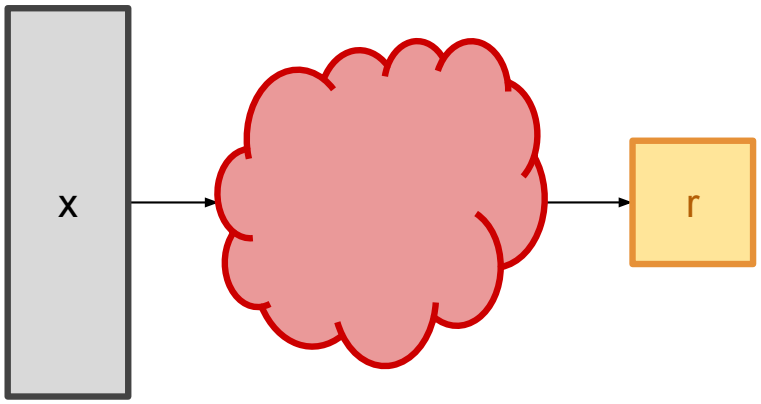
(Representation / Encoder / Inference) Networks



- **Size:** Smaller or larger than x
- **Structure:** Flat or interpretable
- **Type:** Continuous or discrete
- **Shape:** Fixed or variable
- **Disentangled** or not

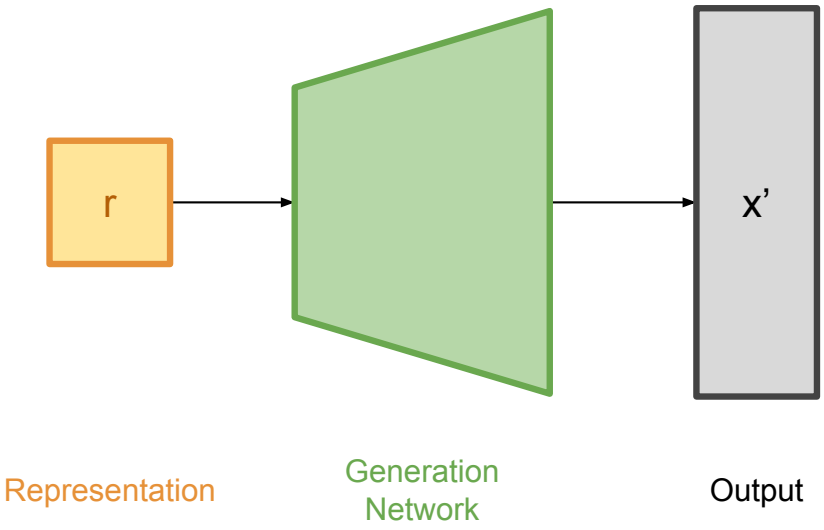
- Multi-layer perceptron
- ConvNet
- Transformer
- Recurrent neural net

(Representation / Encoder / Inference) Networks



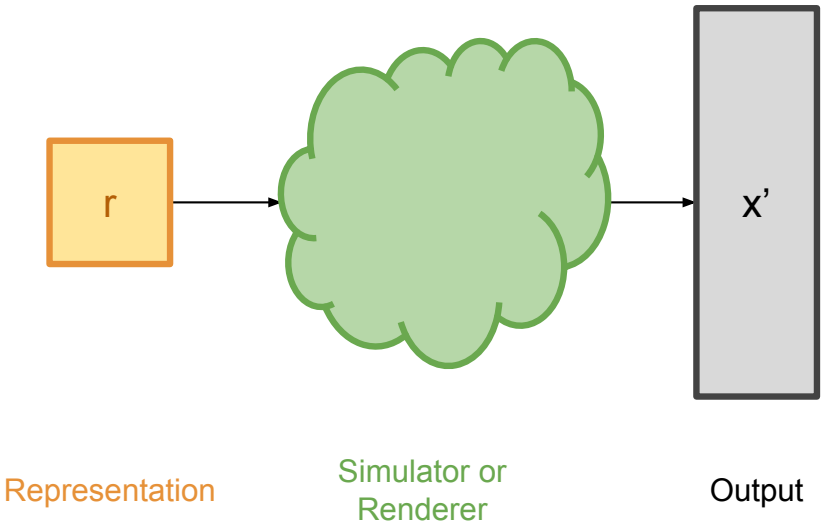
- Differentiable or not
- Interpretable or not
- Deterministic or stochastic

(Generation / Generator / Decoder) Networks



- Multi-layer perceptron
- DeconvNet
- Transformer
- Recurrent neural net

(Generation / Generator / Decoder) Networks



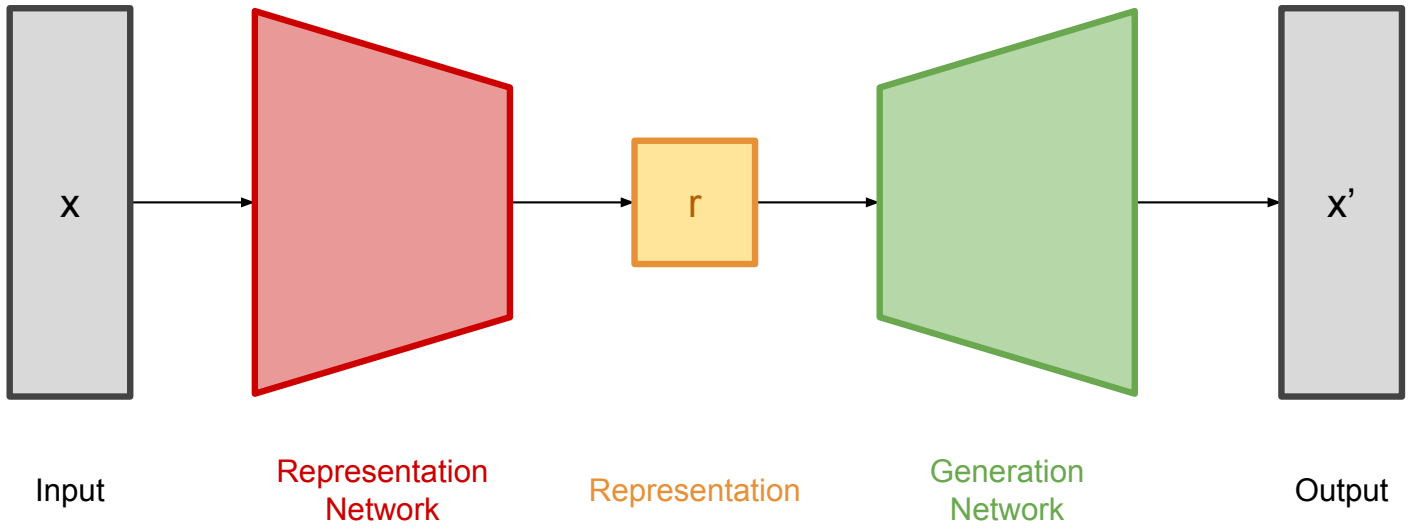
- Differentiable or not
- Interpretable or not
- Deterministic or stochastic

Autoencoders

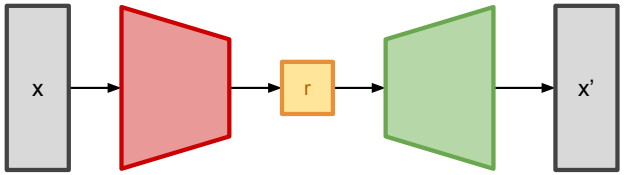
Want to learn more?

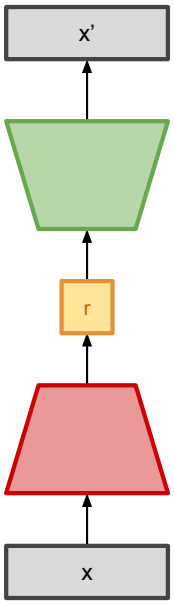
Auto-Encoding Variational Bayes, Kingma et al, ICLR (2014)

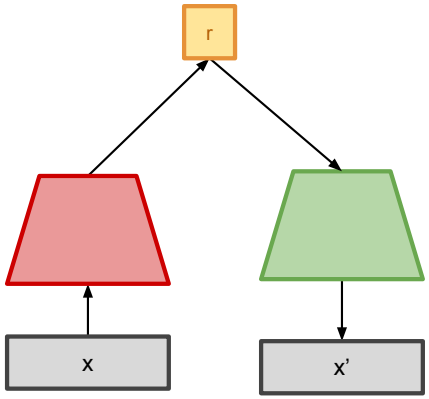
Stochastic backpropagation and approximate inference in deep generative models, Rezende et al, ICML (2014)



Autoencoder Graphics





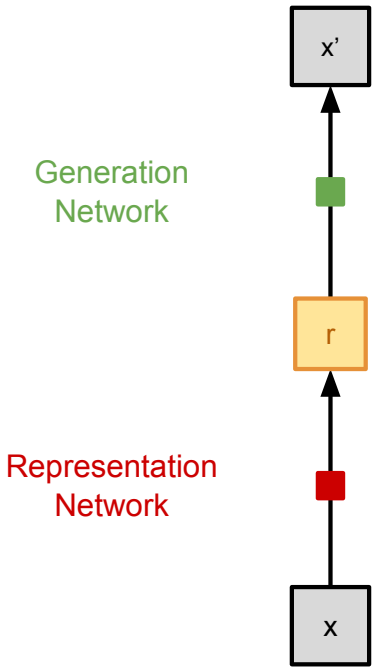


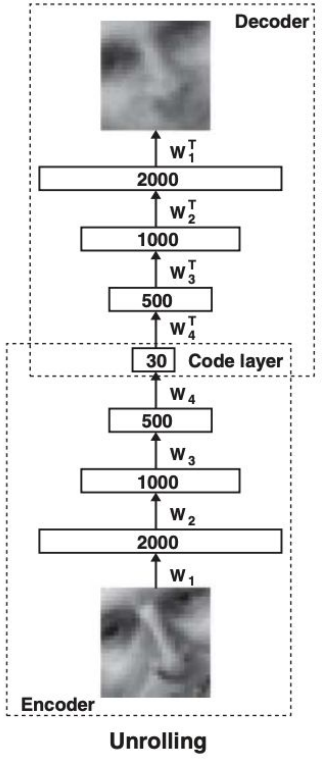
Autoencoders: What are they for?

- Density estimation
- Dimensionality reduction
- Image generation
- Denoising
- **Representation learning**

Want to learn more?

Reducing the Dimensionality of Data with Neural Networks, Hinton et al, Science (2006)



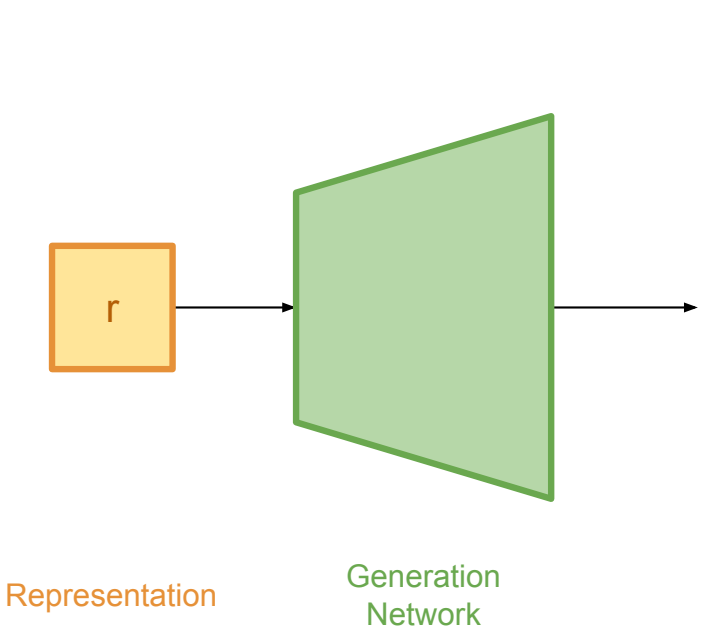


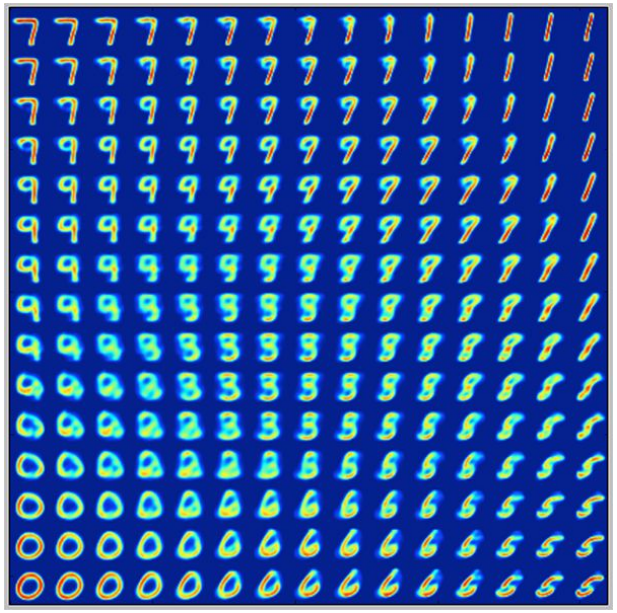
Unrolling

Autoencoders

Want to learn more?

Building Autoencoders in Keras, Chollet (2016)

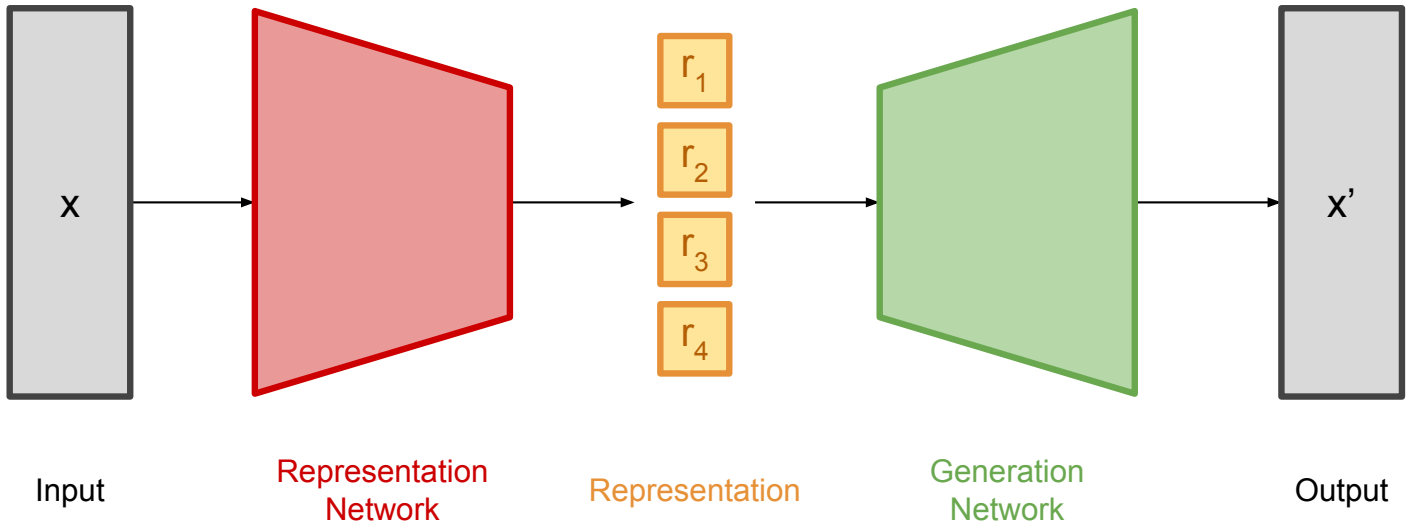




Disentangled Autoencoders

Want to learn more?

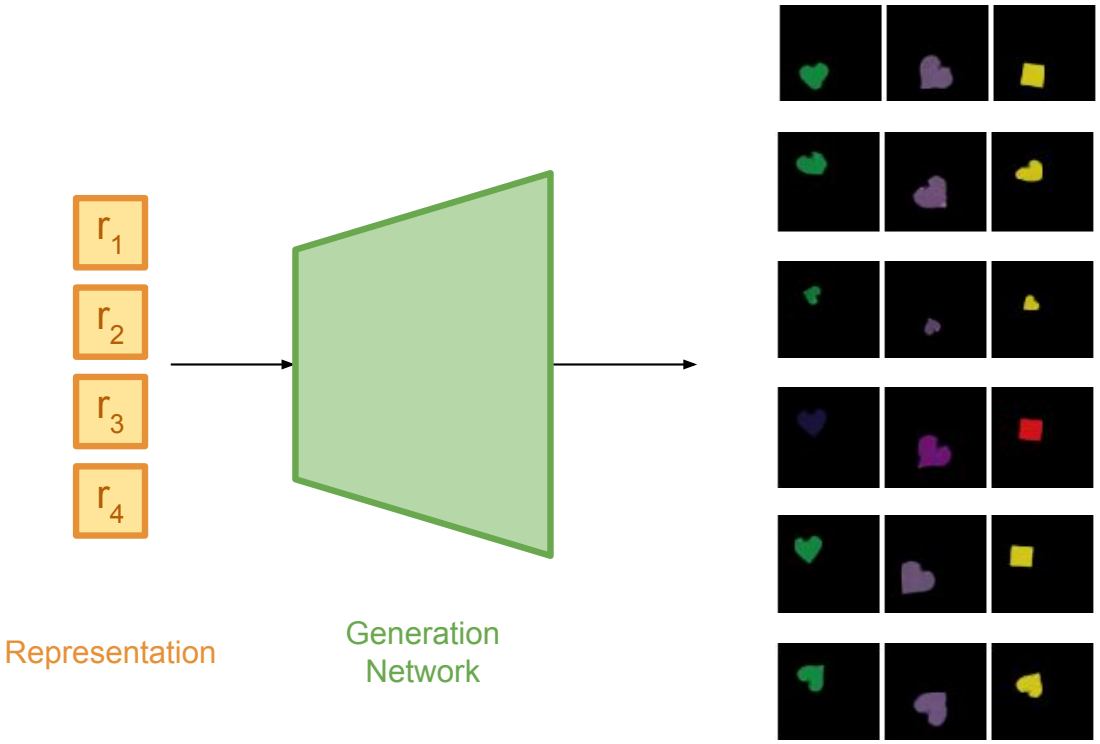
β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, Higgins et al., ICLR 2017



Disentangled Autoencoders

Want to learn more?

β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, Higgins et al., ICLR 2017

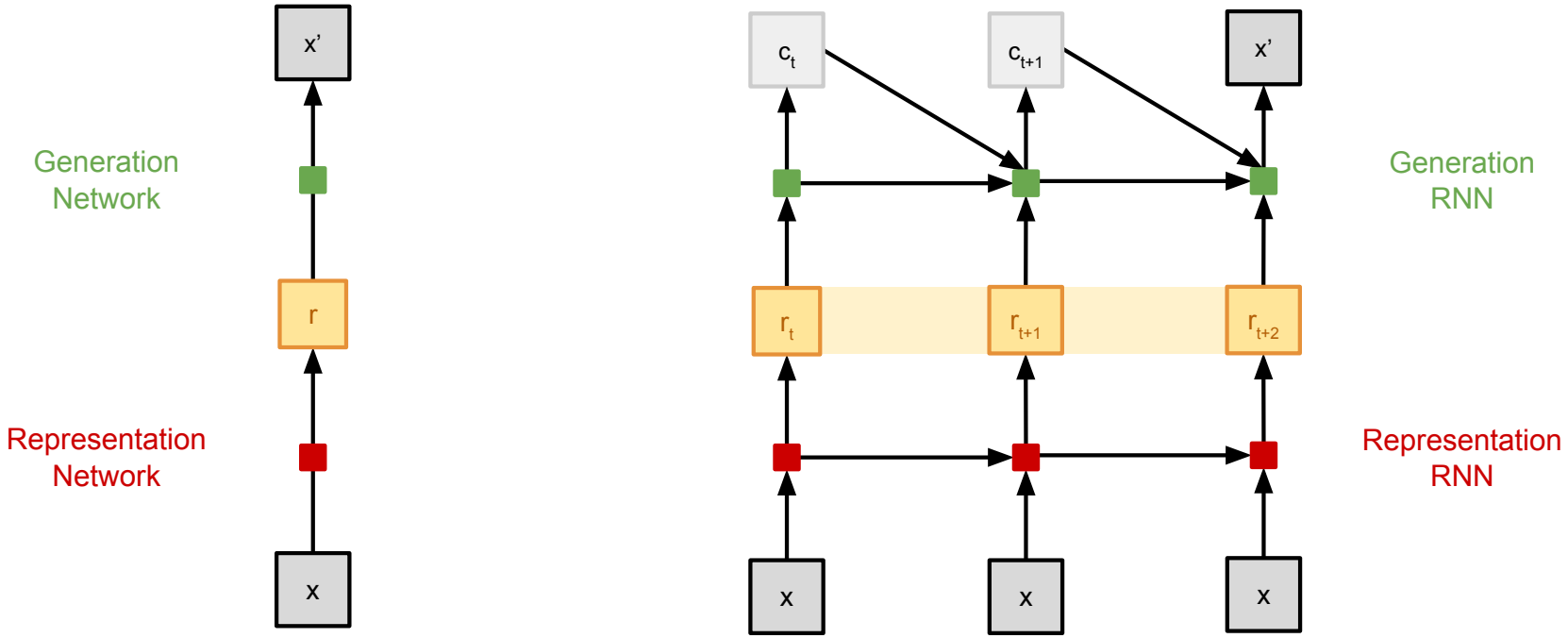


GIFs adapted from Chris Burgess

Sequential Autoencoders

Want to learn more?

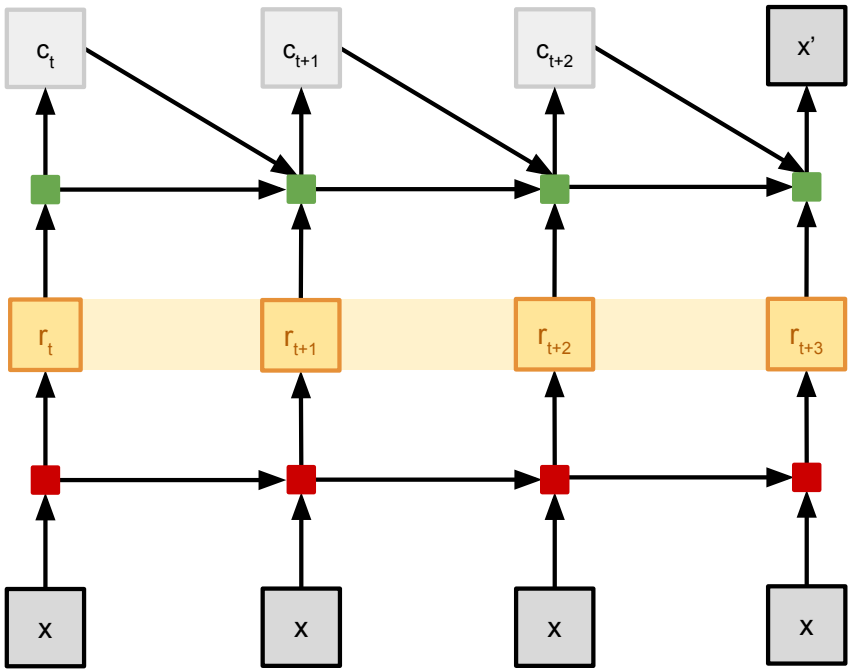
Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



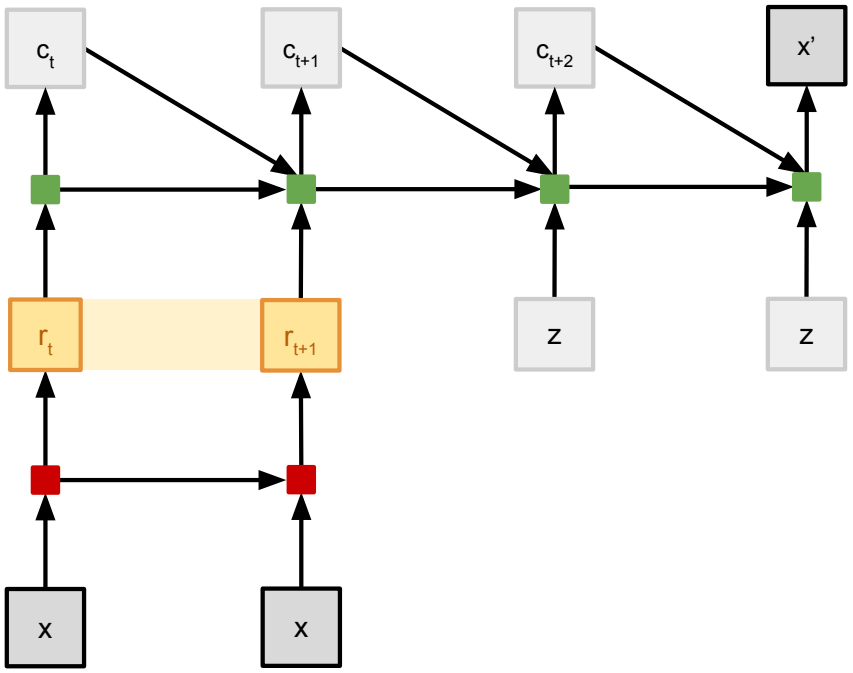
Generation RNN

Representation RNN

Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



Generation RNN

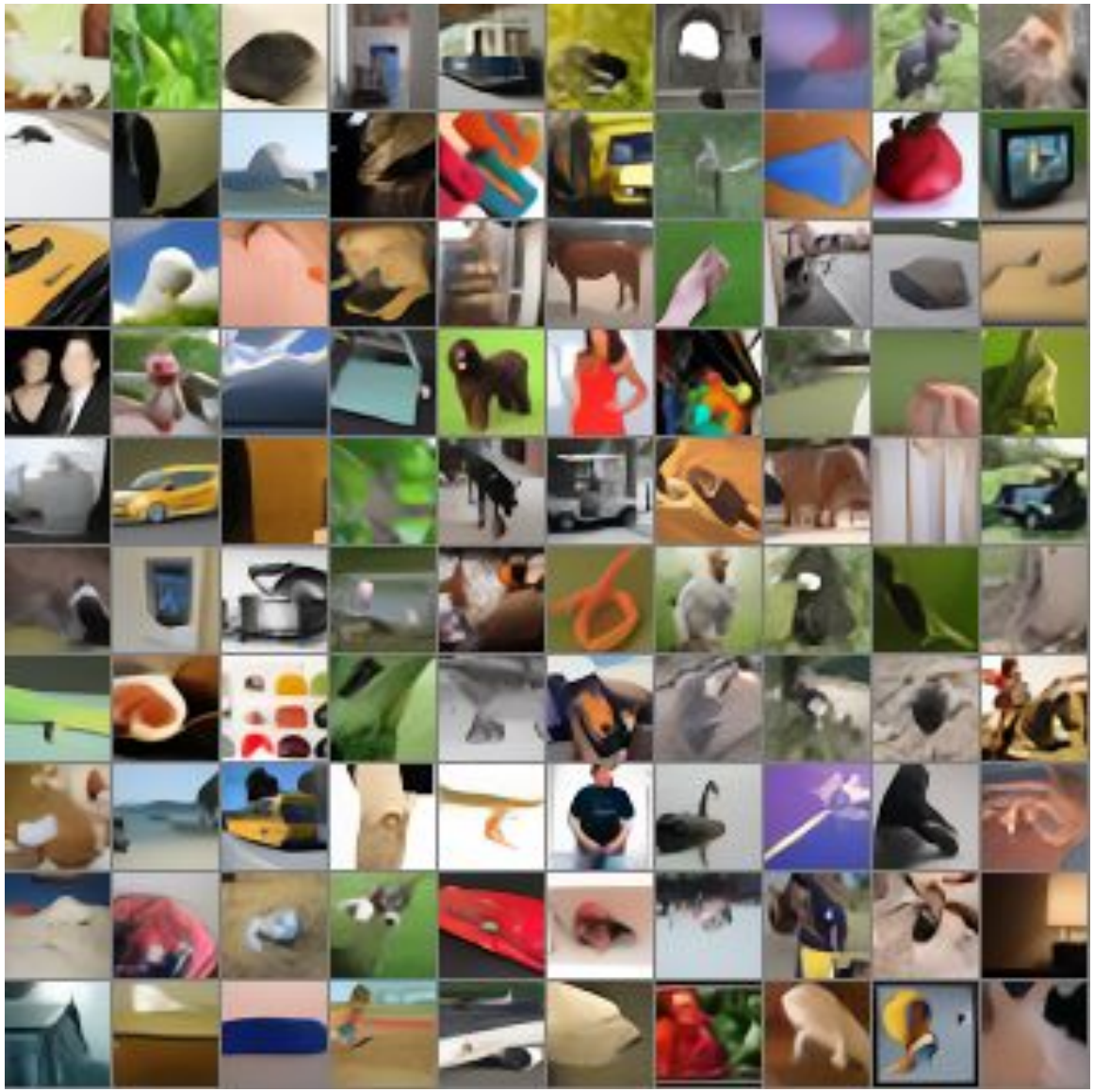
Representation RNN

76 bits

Original raw image:
24576 bits

112 bits

221 bits



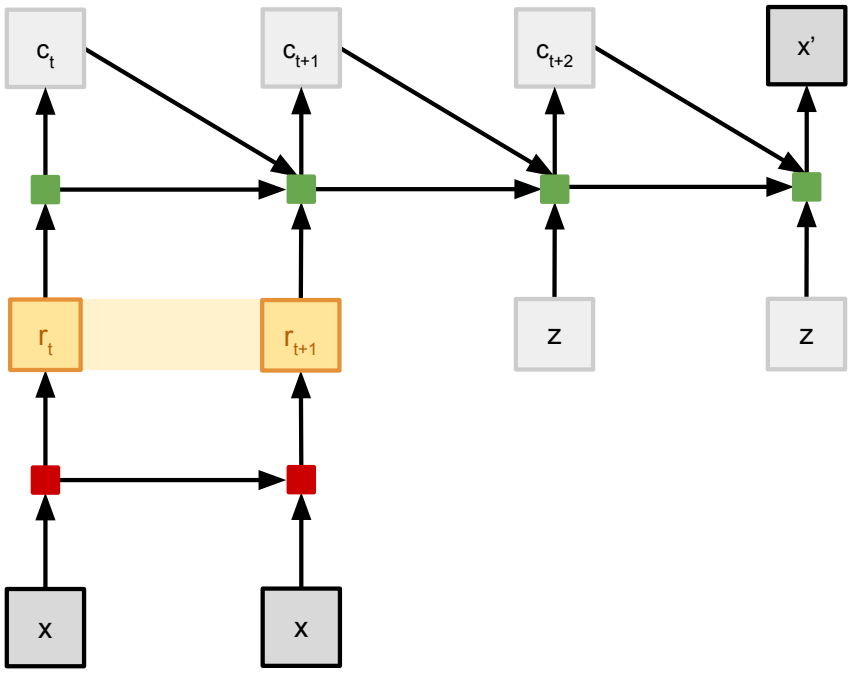
380 bits

2364 bits

Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



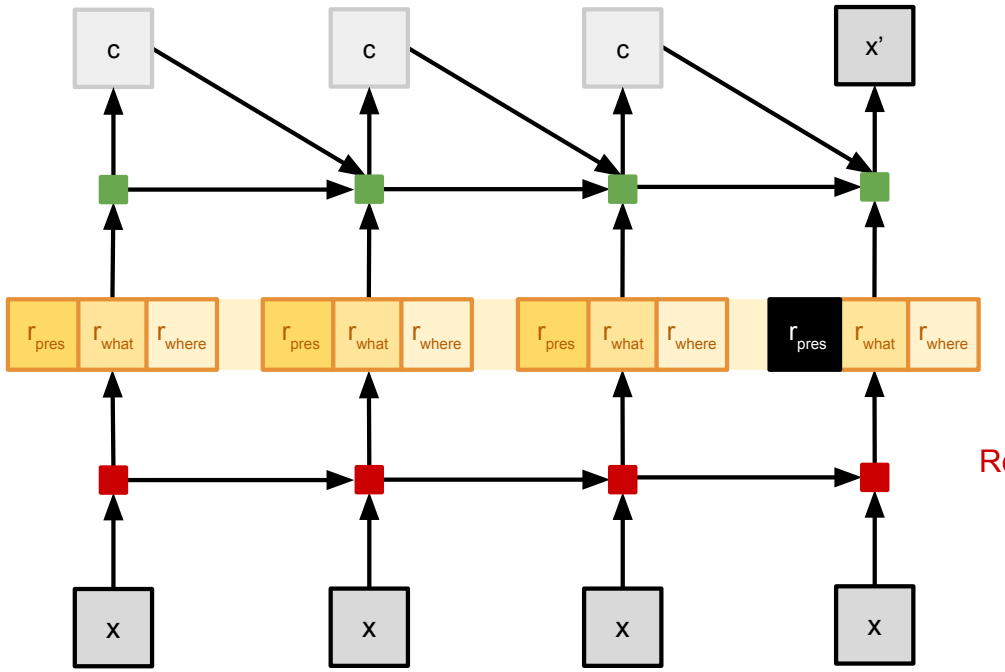
Generation RNN

Representation RNN

Variable Length, Interpretable, Sequential Autoencoders

Want to learn more?

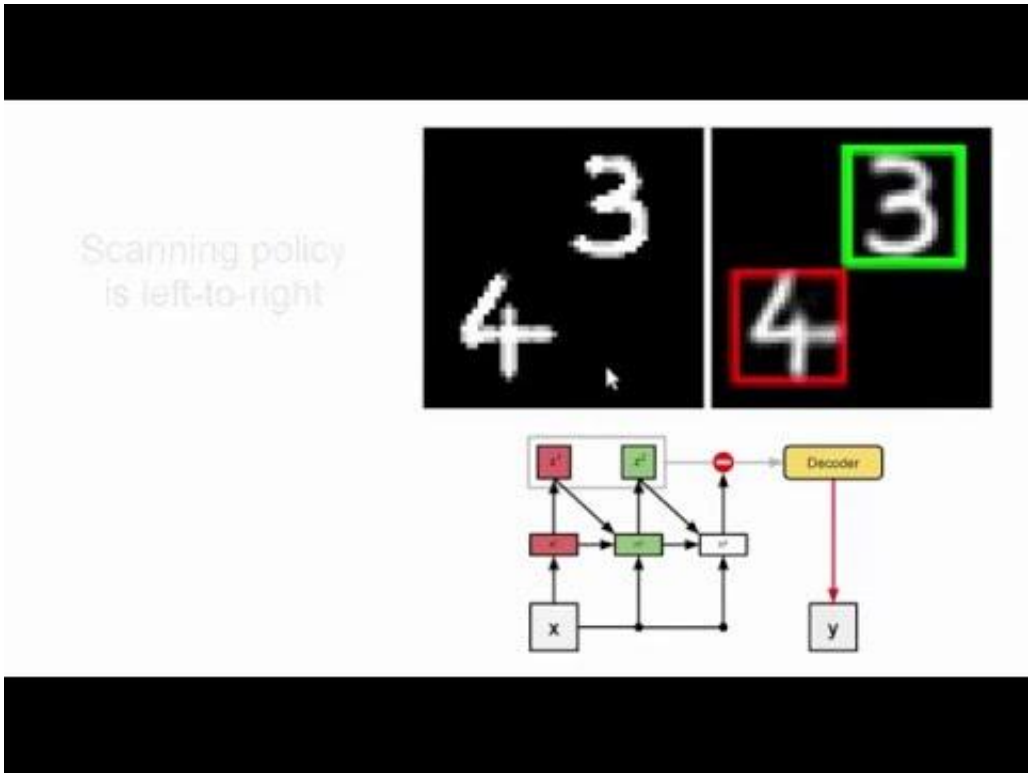
Attend, Infer, Repeat, Eslami et al, NeurIPS (2016)



Variable Length, Interpretable, Sequential Autoencoders

Want to learn more?

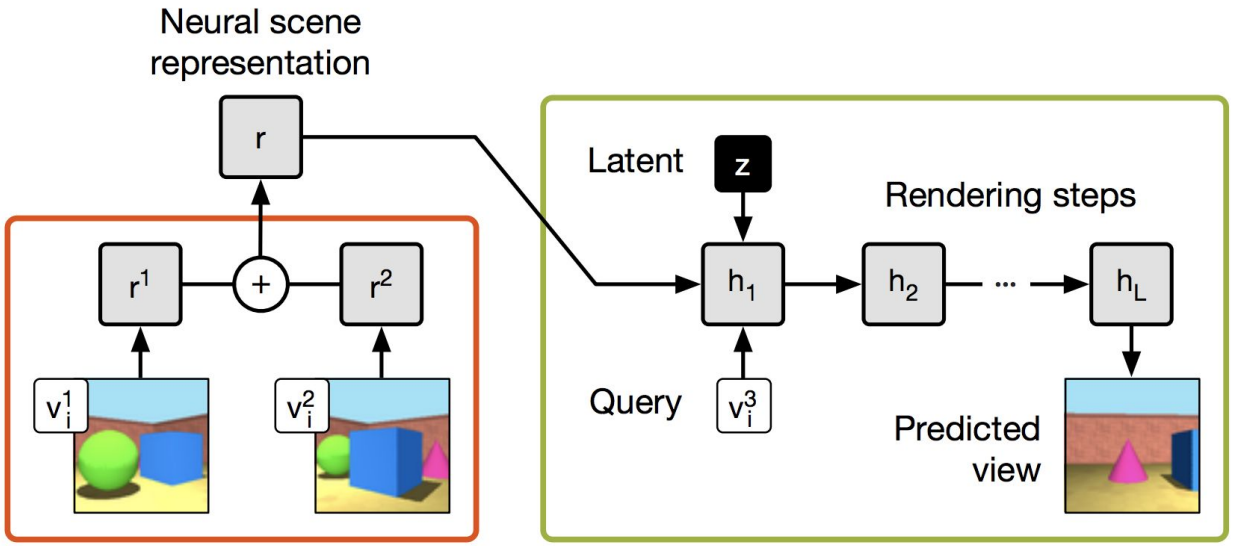
Attend, Infer, Repeat, Eslami et al, NeurIPS (2016)



Generative Query Networks

Want to learn more?

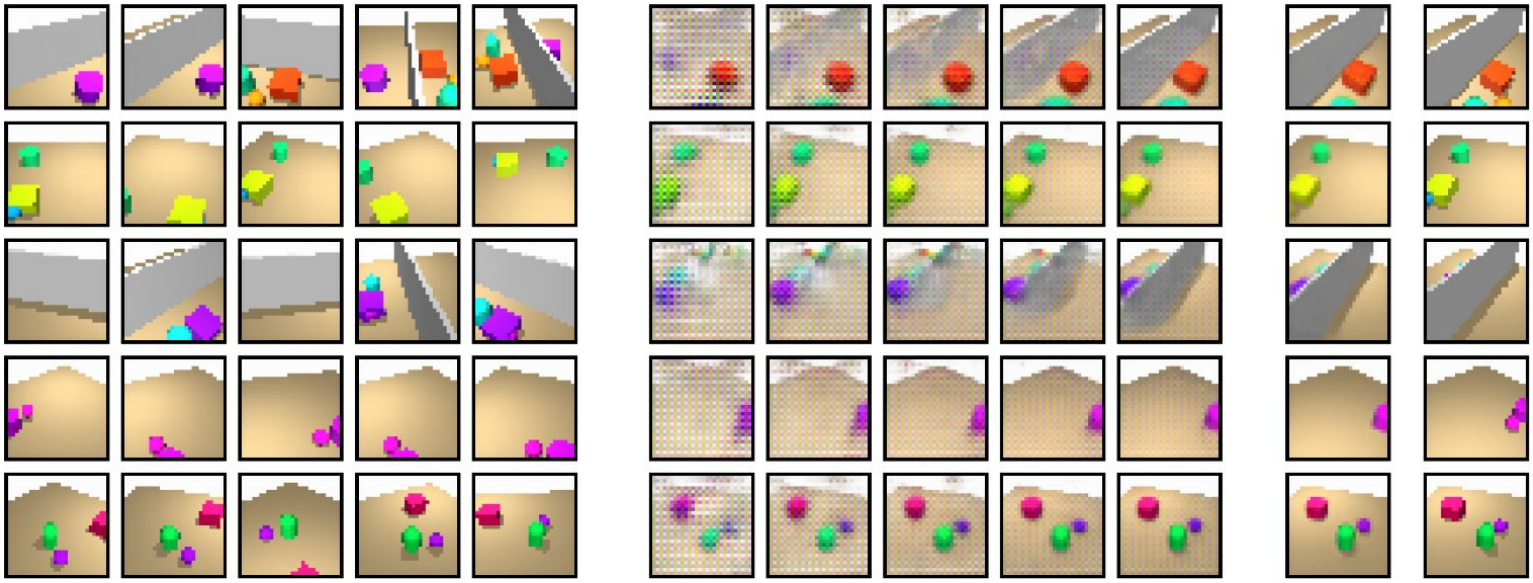
Neural scene representation and rendering, Eslami et al, Science (2018)



GQN: Accurate generation

Want to learn more?

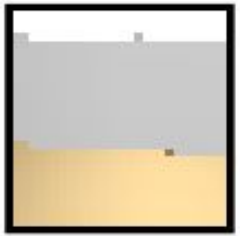
Neural scene representation and rendering, Eslami et al, Science (2018)



GQN: Capturing uncertainty

Want to learn more?

Neural scene representation and rendering, Eslami et al, Science (2018)



Observation

Samples

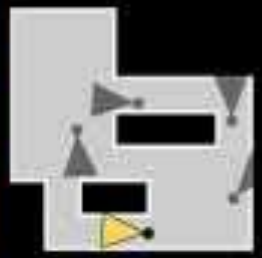
observations

ground truth

neural rendering

neural rendering

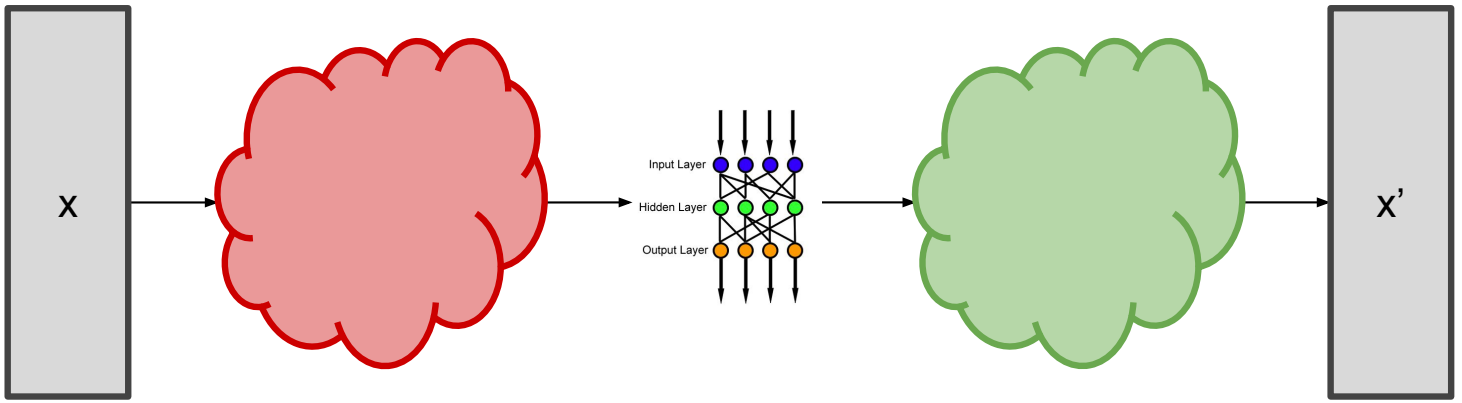
map



NeRF

Want to learn more?

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall (2020)



Input

Inference Optimisation Algorithm

Neural Network Representation

Forward Propagation

Output

NeRF

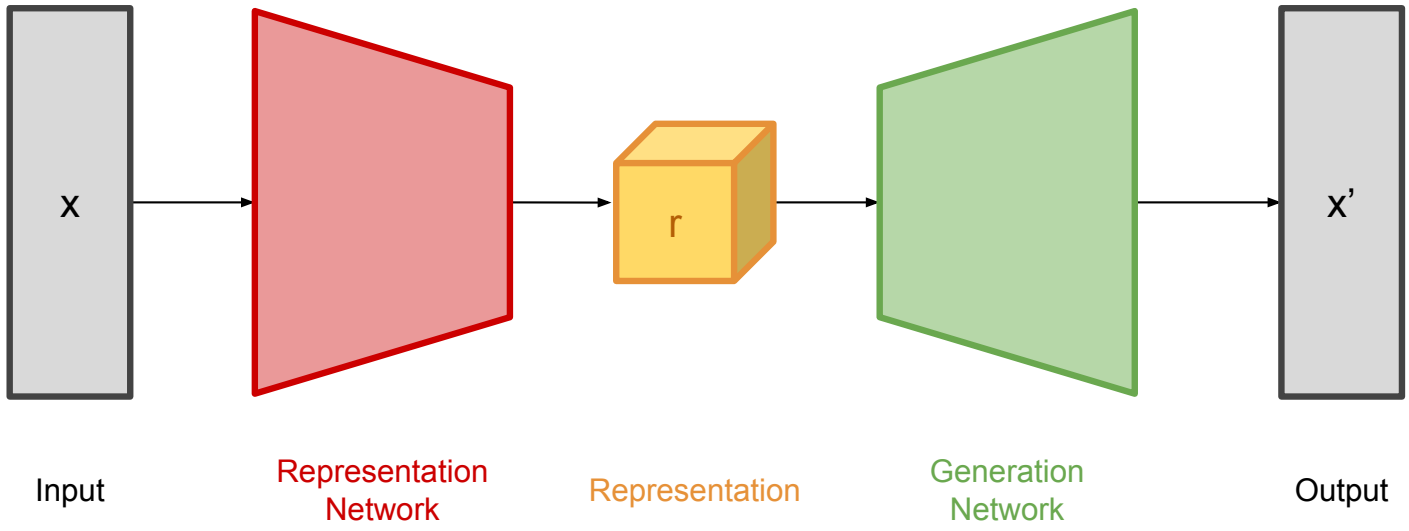
Want to learn more?

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall (2020)

Voxel Autoencoders

Want to learn more?

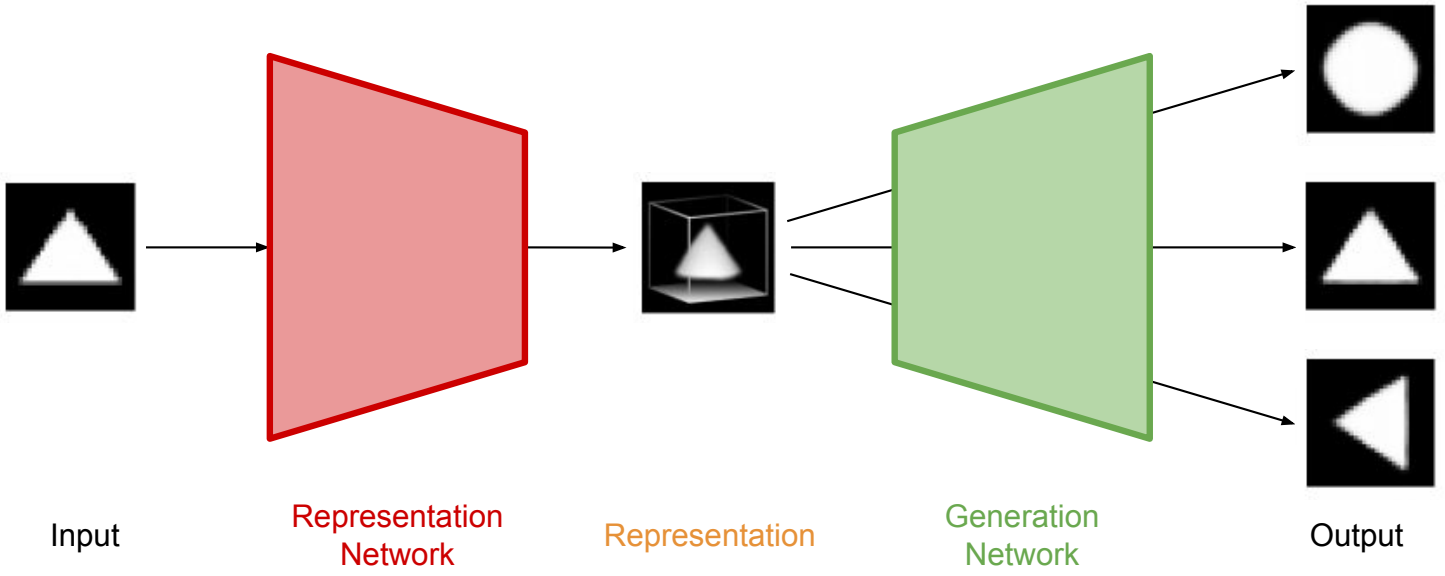
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Voxel Autoencoders

Want to learn more?

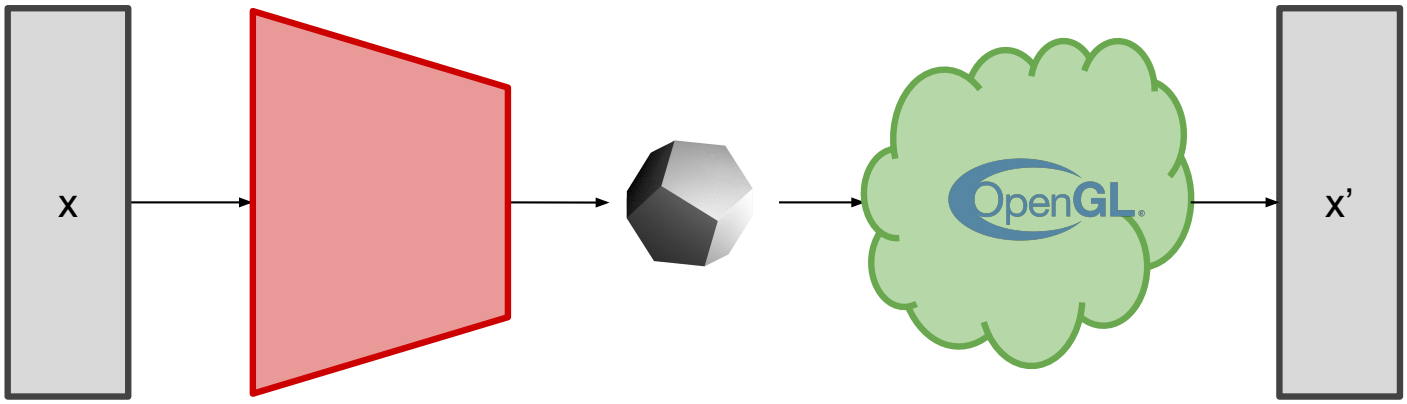
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Mesh Autoencoders

Want to learn more?

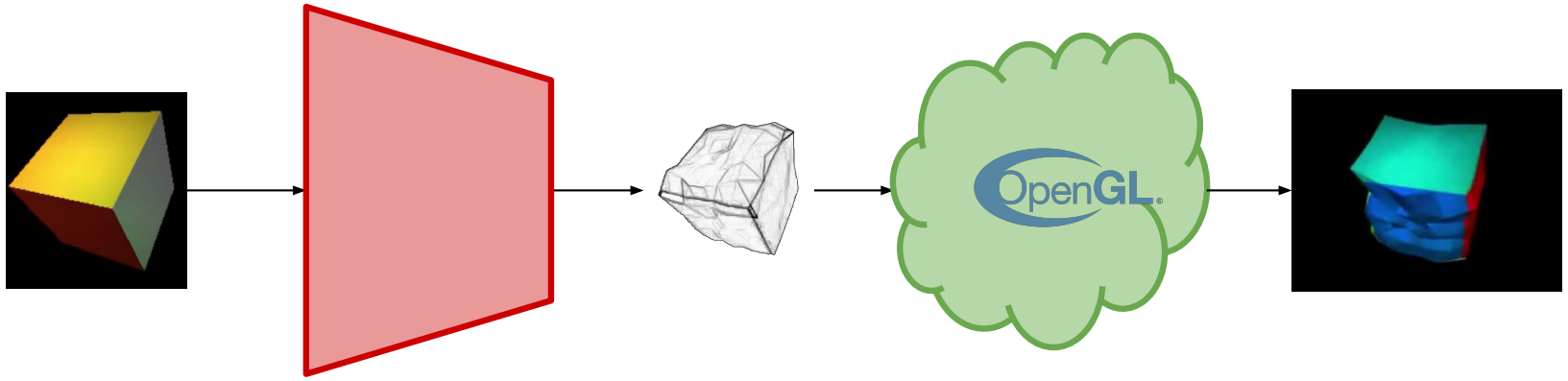
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Mesh Autoencoders

Want to learn more?

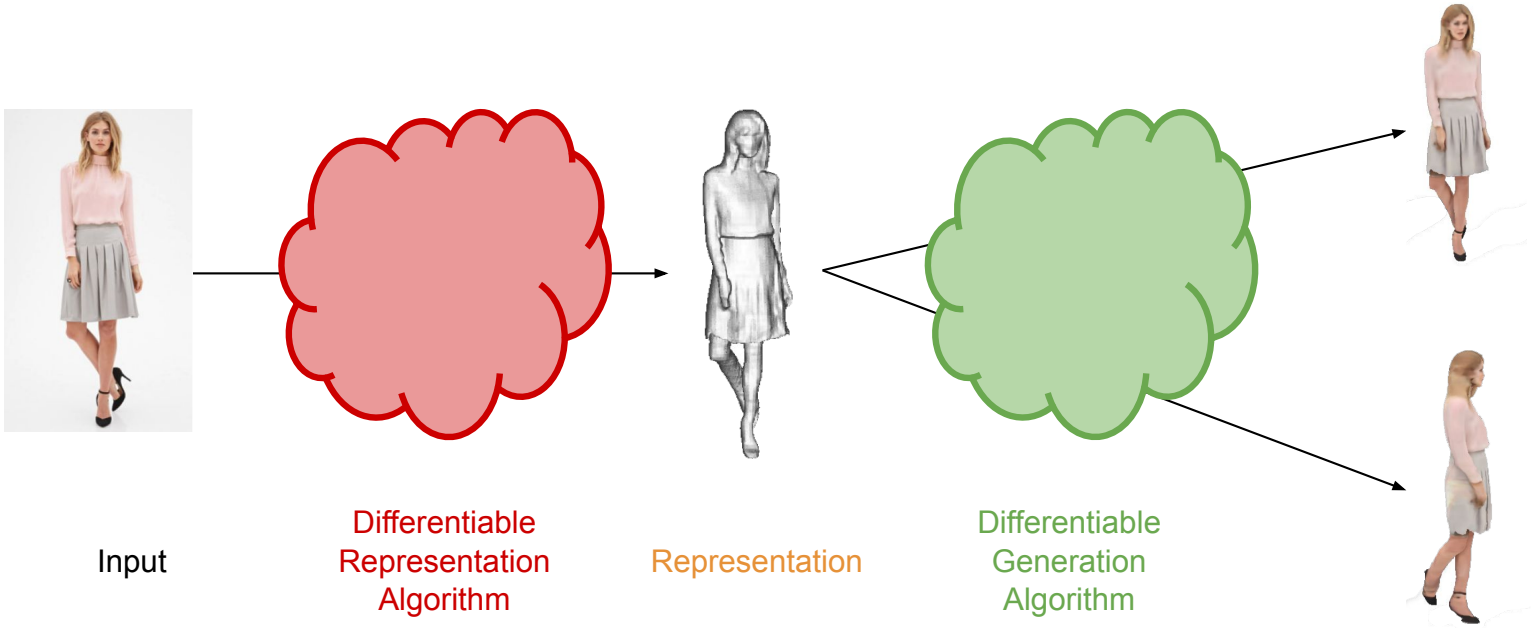
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Implicit Function Autoencoders

Want to learn more?

PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization, Saito et al, (2019)



Beyond likelihood-based

Discriminators / Contrastive Networks

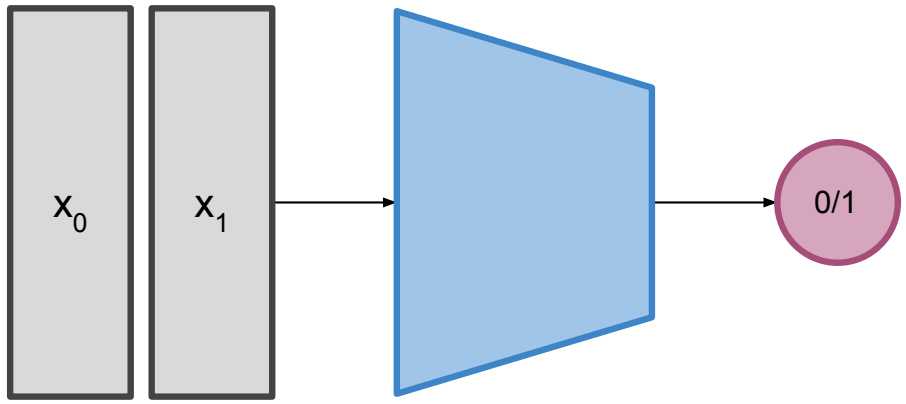


Diagram inspired by Lilian Weng, https://lilianweng.github.io

Generative adversarial networks

Want to learn more?

Generative adversarial networks.
Goodfellow, et al. NeurIPS (2014)

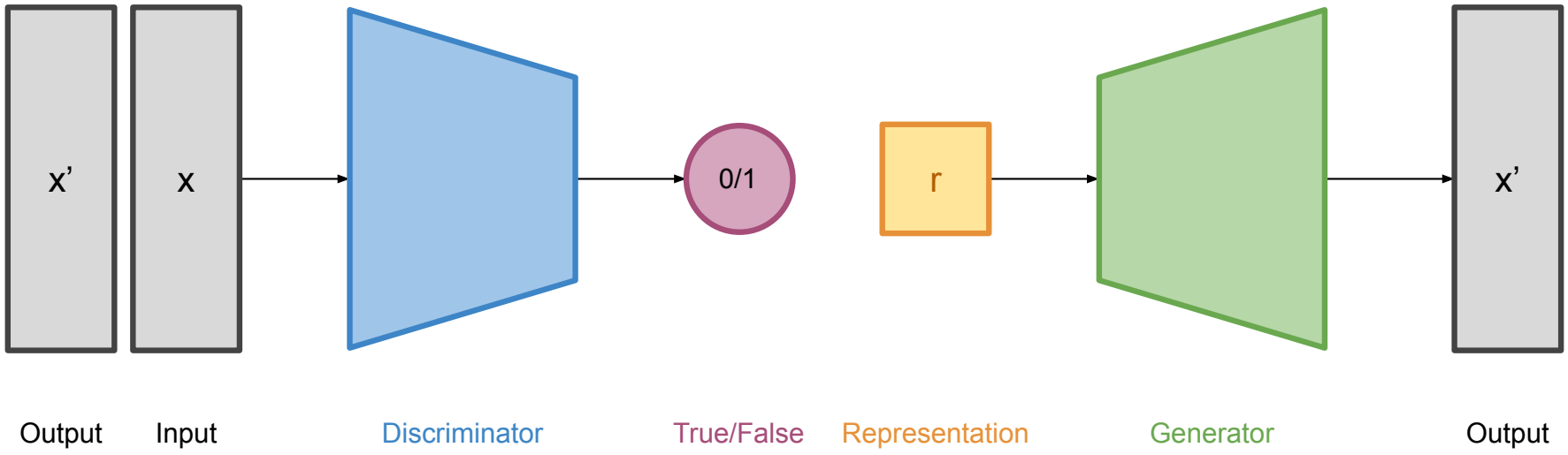


Diagram inspired by Lilian Weng, https://lilianweng.github.io

Generative adversarial networks

Want to learn more?

A Style-Based Generator for GANs, Karras et al (2018)

Large Scale GAN Training for High Fidelity Natural Image Synthesis, Brock et al (2018)

Generative adversarial networks

Want to learn more?

Generative adversarial networks.
Goodfellow, et al. NeurIPS (2014)

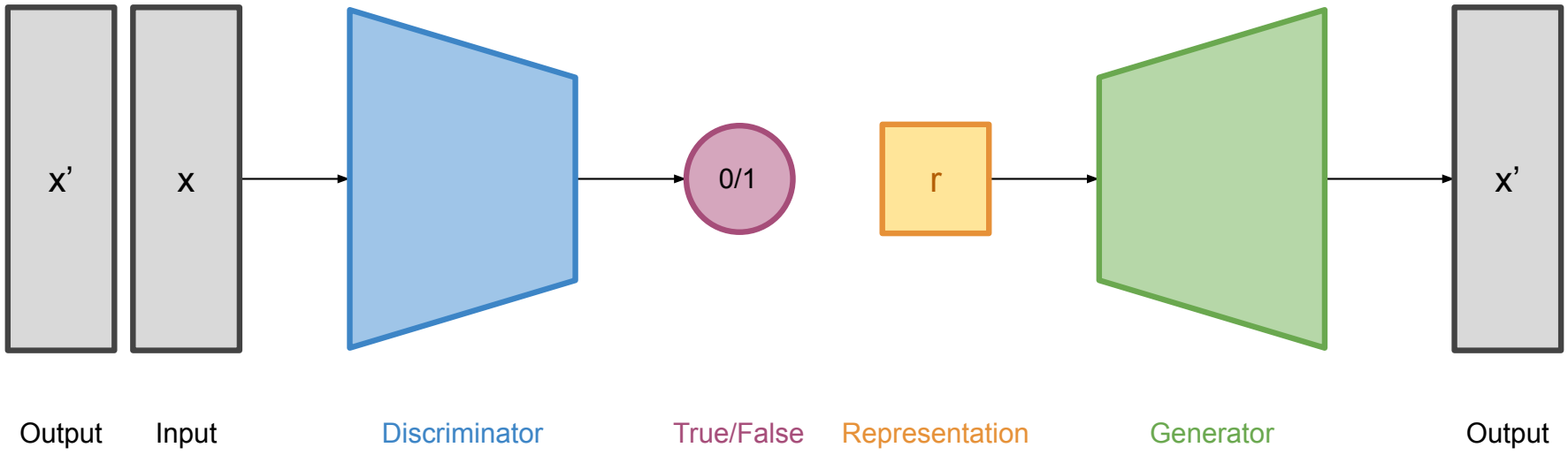
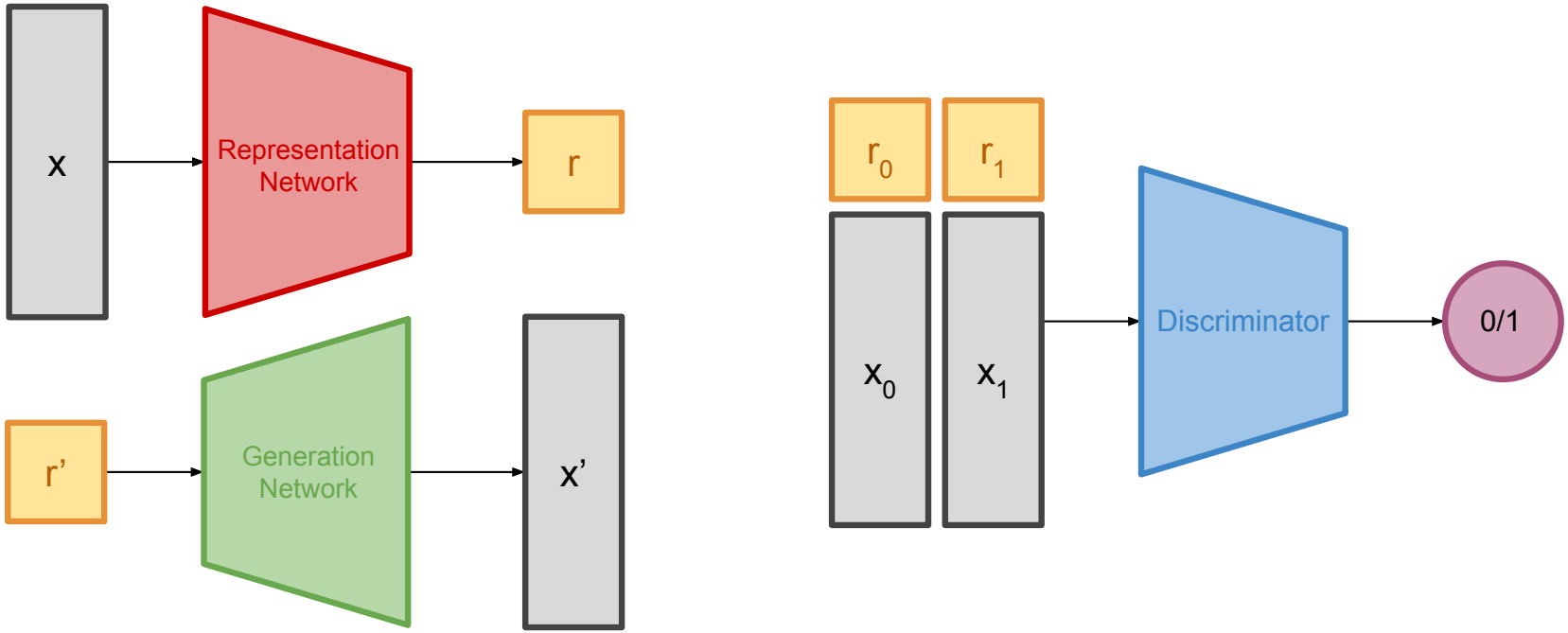


Diagram inspired by Lilian Weng, https://lilianweng.github.io

BiGAN

Want to learn more?

Adversarial Feature Learning, Donahue, et al. ICLR (2017)



BigBiGAN

Want to learn more?

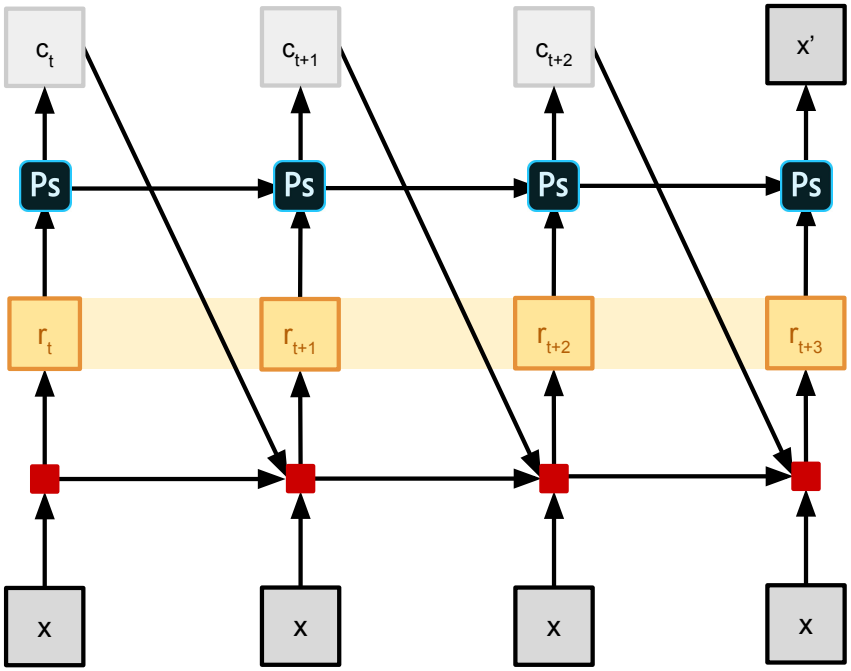
Large Scale Adversarial Representation Learning. Donahue, et al. NeurIPS (2019)

SPIRAL

Want to learn more?

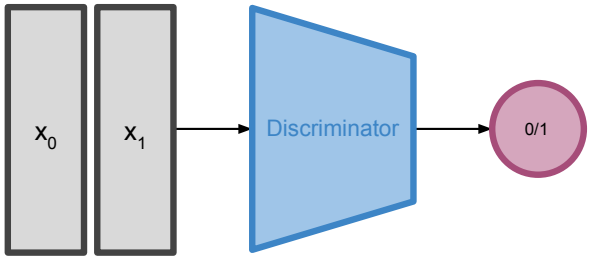
Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

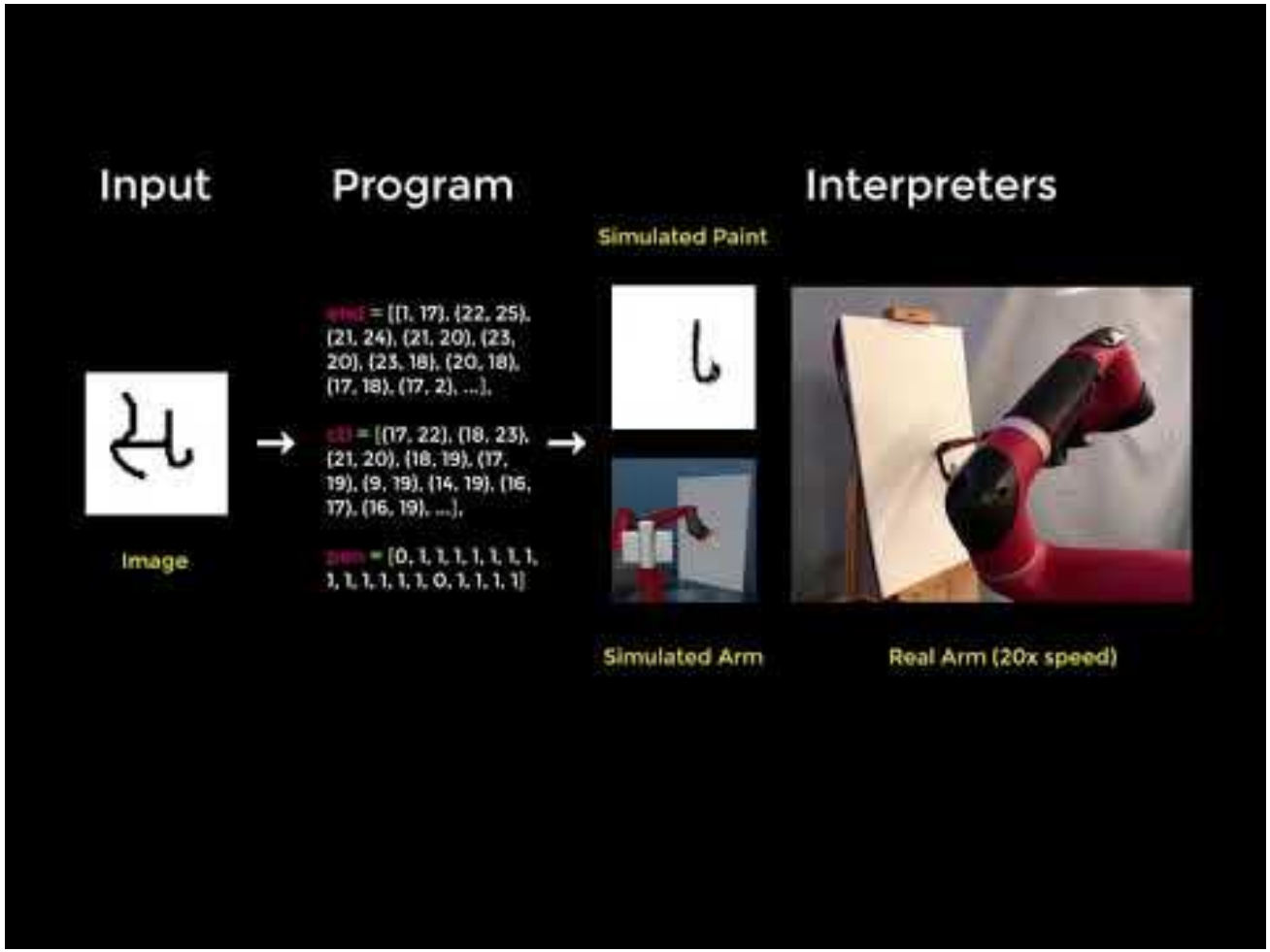
Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)



Paintbrush Simulator

Representation RNN





Want to learn more?

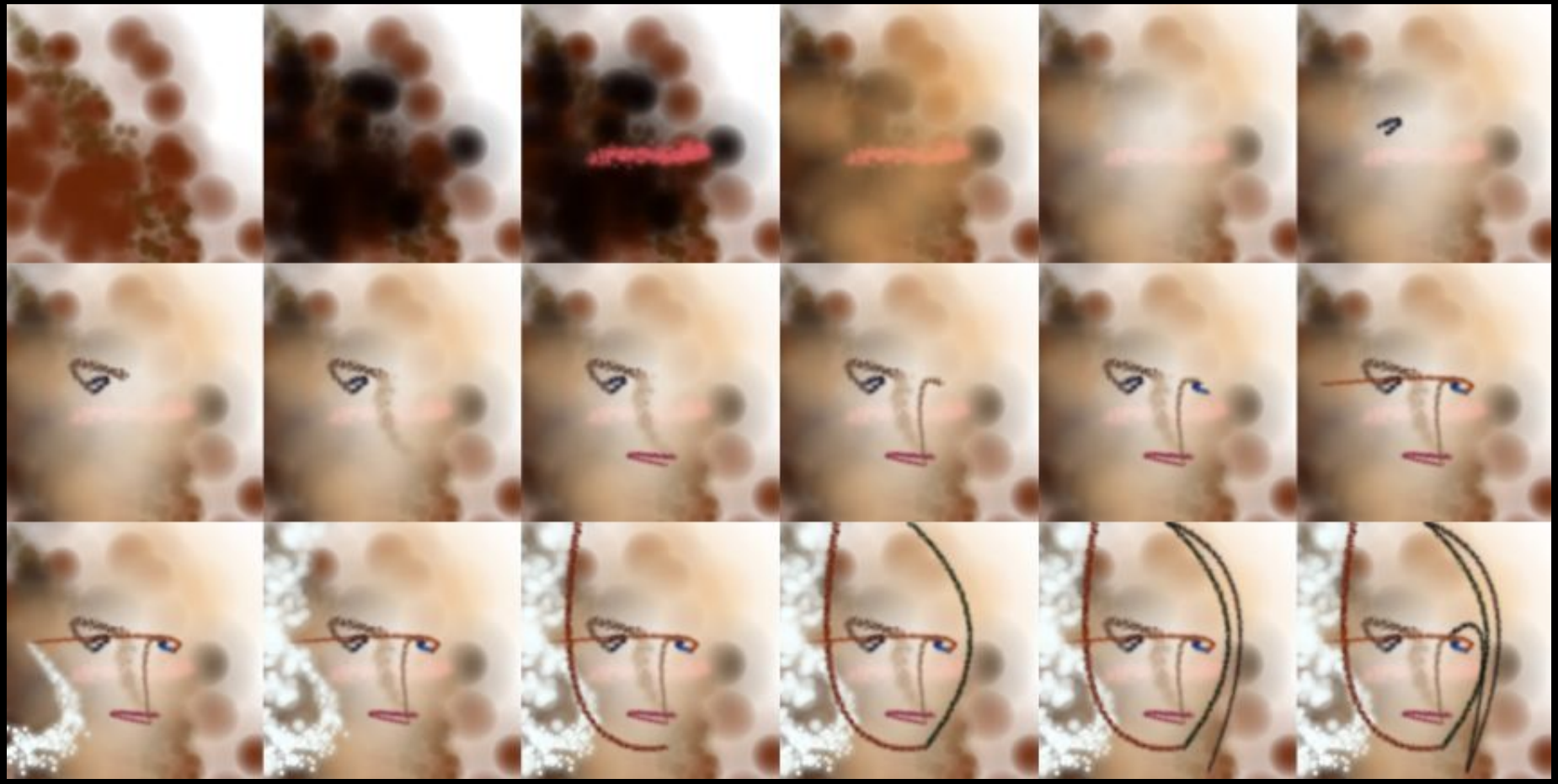
Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)

Want to learn more?

Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)

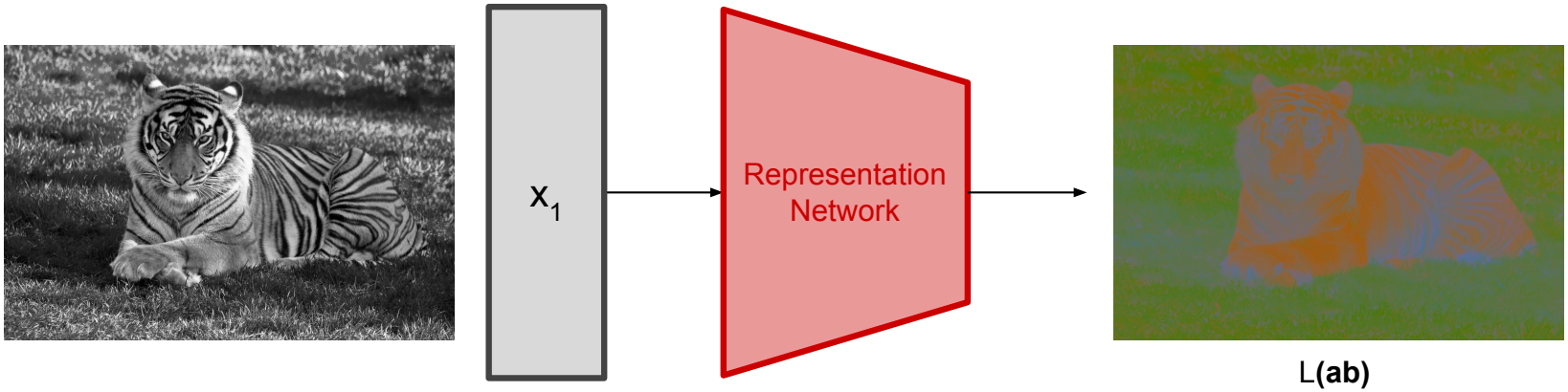


Beyond generative

Colorization

Want to learn more?

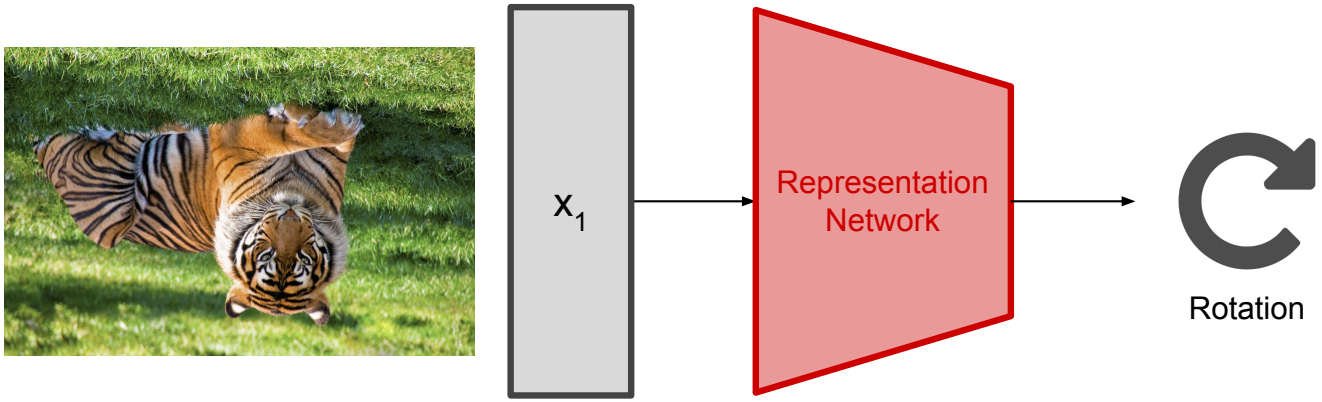
Colorization as a proxy task for visual understanding, Larsson et al, CVPR (2017)



Rotation Prediction

Want to learn more?

Unsupervised Representation Learning by Predicting Image Rotations, Gidaris et al, ICLR (2018)

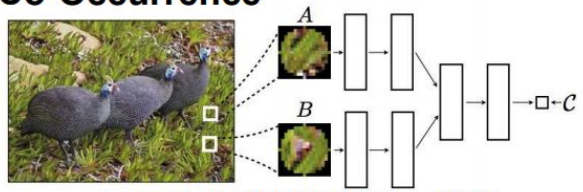


Self-supervised learning

Want to learn more?

Self-Supervised Learning lecture, Andrew Zisserman, ICML (2018)

Co-Occurrence

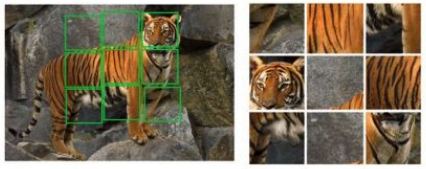


Isola *et al.* ICLR Workshop 2016.

Egomotion

Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

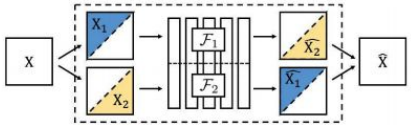
Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders

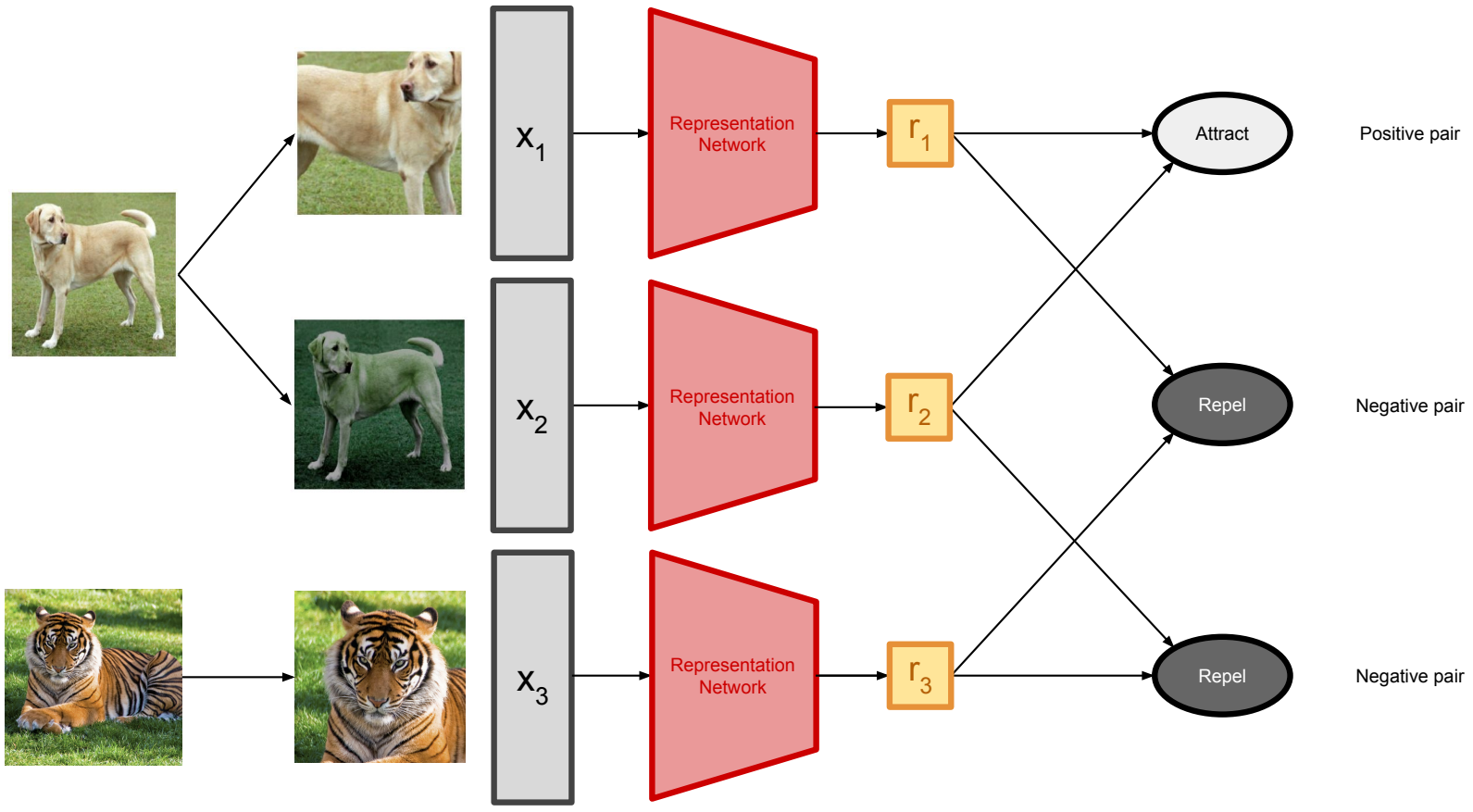


Zhang *et al.* CVPR 2017

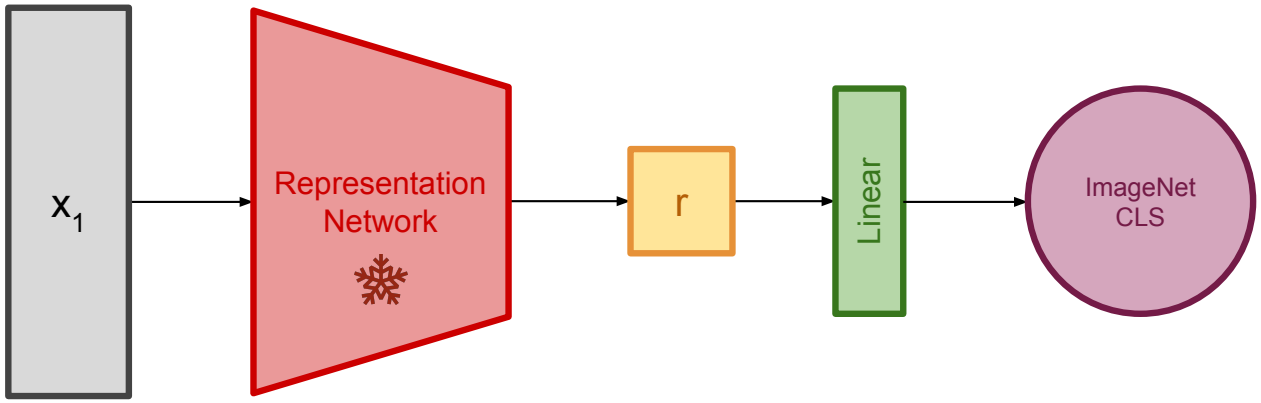
Contrastive learning

Want to learn more?

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML (2020)



Evaluation: Linear separation



Contrastive learning

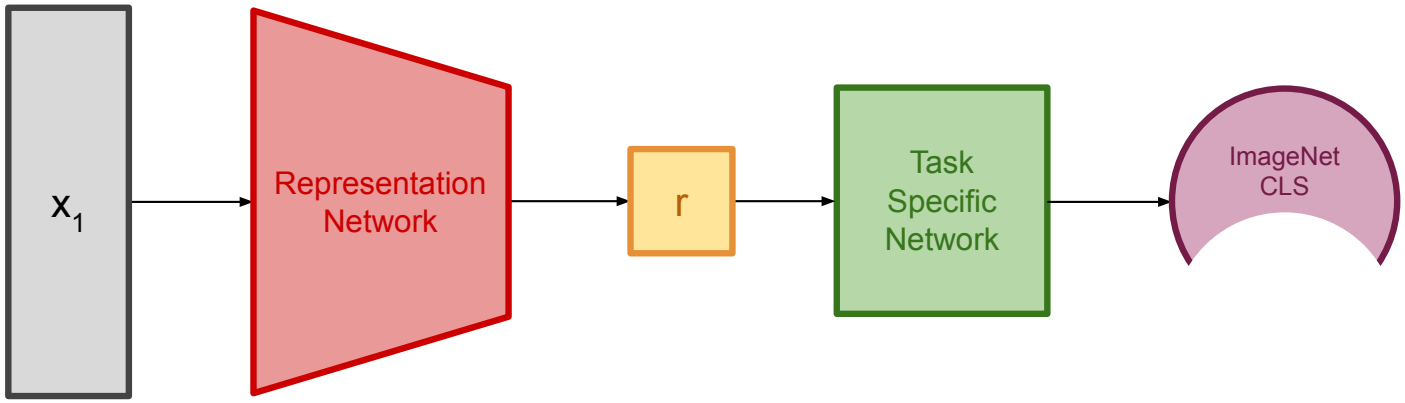
Want to learn more?

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML (2020)

Method	Architecture	Param.	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (∗)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

Evaluation: Data efficiency



Data efficient representation learning

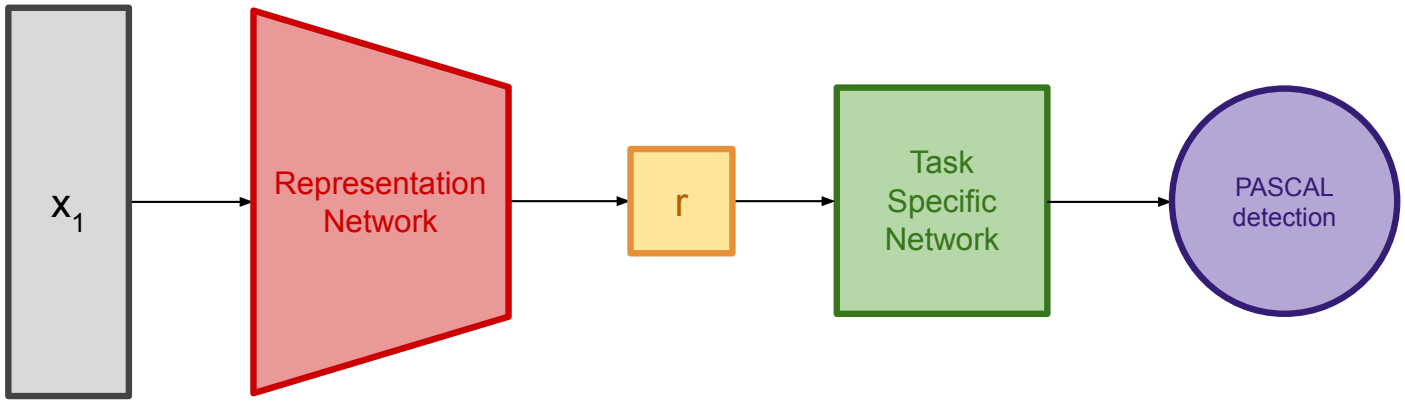
Want to learn more?

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML (2020)

Method	Architecture	Label fraction 1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Evaluation: Transfer learning



Transfer learning

Want to learn more?

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML (2020)

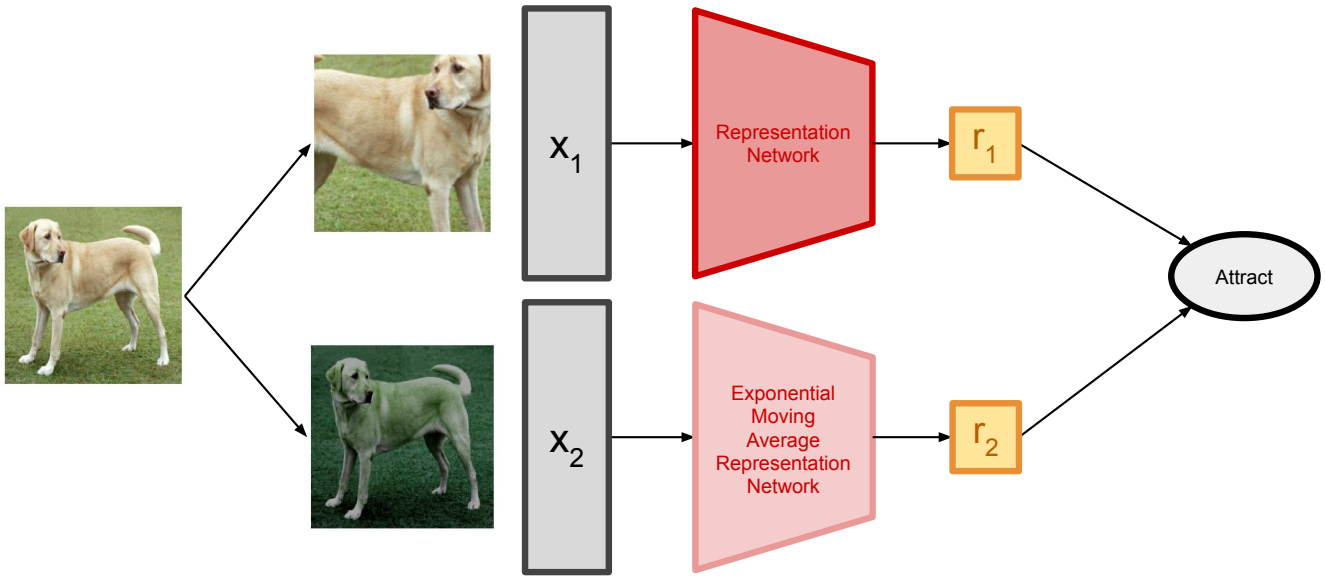
	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

Bootstrap Your Own Latent

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv (2020)



Bootstrap Your Own Latent

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv (2020)

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [29]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL	ResNet-50 (4×)	375M	78.6	94.2
BYOL	ResNet-200 (2×)	250M	**79.6**	**94.8**

Surprising return of likelihood-based models

Attention-based models

iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)

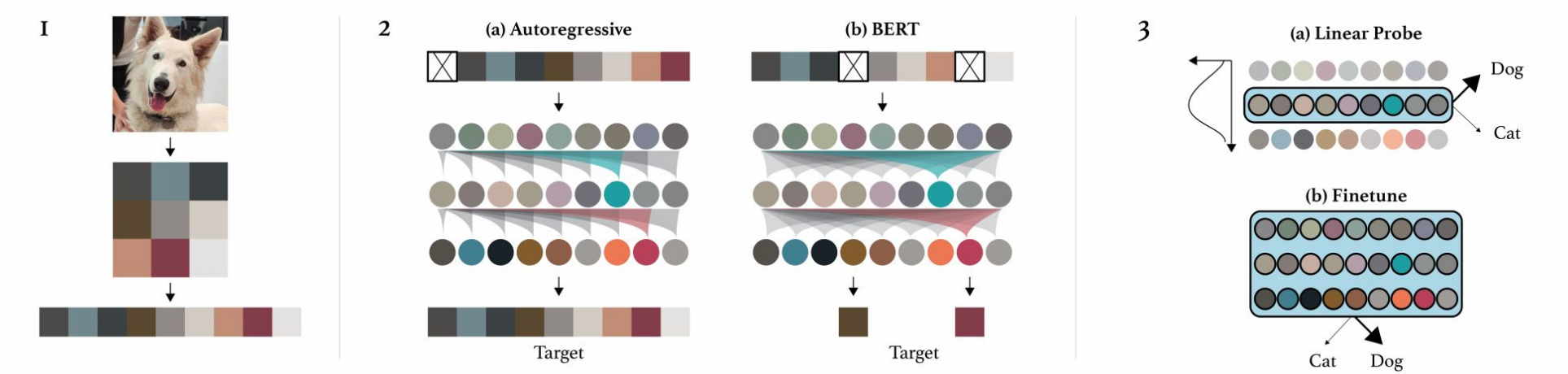
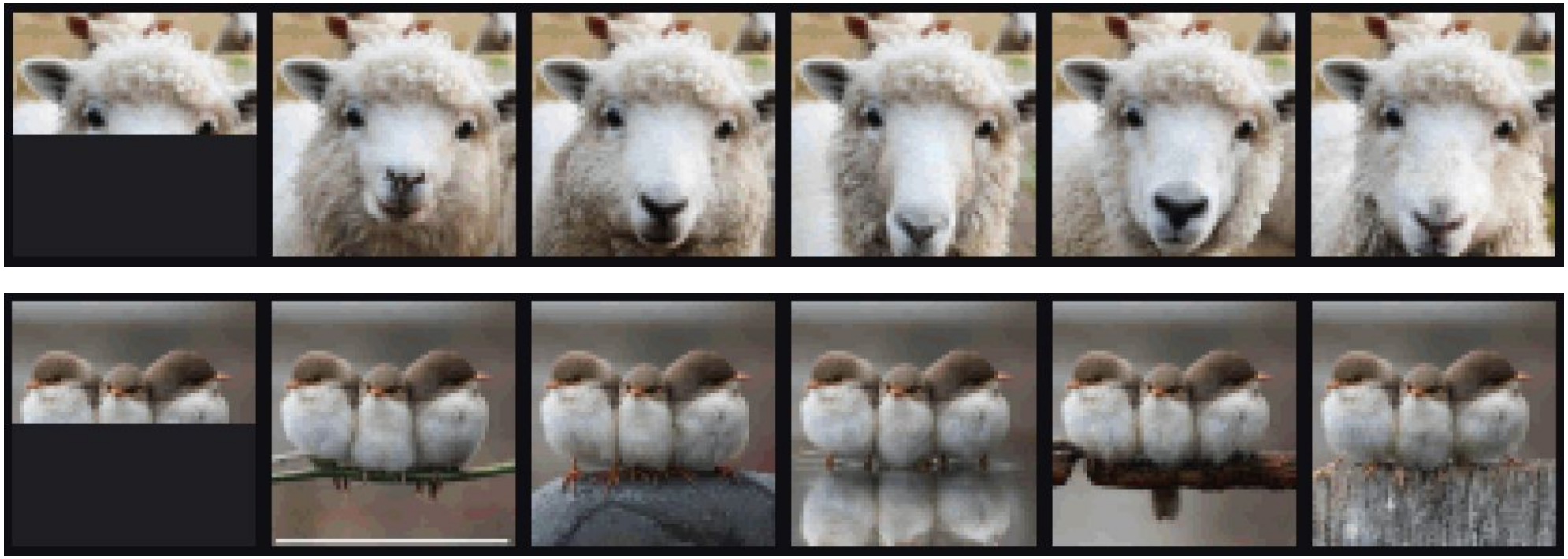


Figure 1. An overview of our approach. First, we pre-process raw images by resizing to a low resolution and reshaping into a 1D sequence. We then chose one of two pre-training objectives, auto-regressive next pixel prediction or masked pixel prediction. Finally, we evaluate the representations learned by these objectives with linear probes or fine-tuning.

iGPT

Want to learn more?

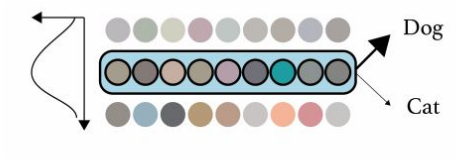
Generative Pretraining from Pixels, Chen et al, ICML (2020)

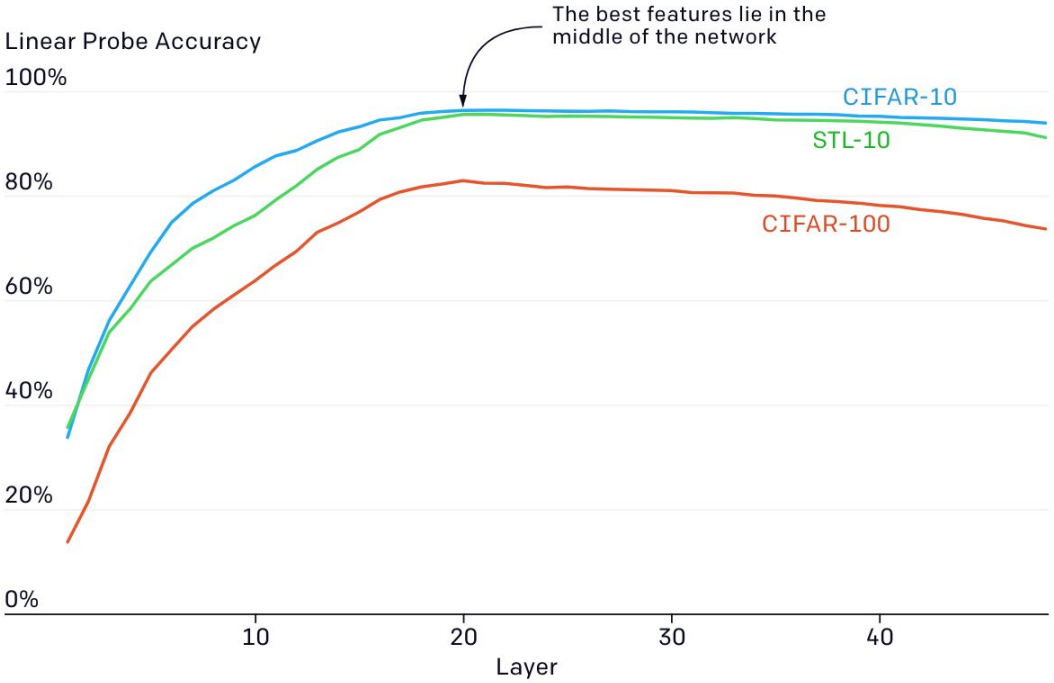


iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)





Feature quality depends heavily on the layer we choose to evaluate. In contrast with supervised models, the best features for these generative models lie in the middle of the network.

iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)

METHOD	INPUT RESOLUTION	FEATURES	PARAMETERS	ACCURACY
Rotation[53]	original	8192	86M	55.4
iGPT-L	32x32	1536	1362M	60.3
BigBiGAN[37]	original	8192	86M	61.3
iGPT-L	48x48	1536	1362M	65.2
AMDIM[13]	original	8192	626M	68.1
MoCo[24]	original	8192	375M	68.6
iGPT-XL	64x64	3072	6801M	68.7
SimCLR[12]	original	2048	24M	69.3
CPC v2[25]	original	8192	303M	71.5
iGPT-XL	64x64	3072 x 5	6801M	72.0
SimCLR	original	8192	375M	**76.5**

A comparison of linear probe accuracies between our models and state-of-the-art self-supervised models. We achieve competitive performance while training at much lower input resolutions, though our method requires more parameters and compute.

Summary

The representation learning problem is **under-specified**.

Two broad categories of approaches:

1. **Building in structure** or **inductive bias** to obtain the 'right' representations, e.g. structured autoencoders
2. **Training for proxy tasks** that can only be solved with the 'right' representations, e.g. contrastive learning

Current approaches seem to involve a trade-off between:

1. **Generality of the representation**, i.e. what range of downstream tasks the representation is good for
2. **Interpretability**, i.e. how much control we have on the representational space

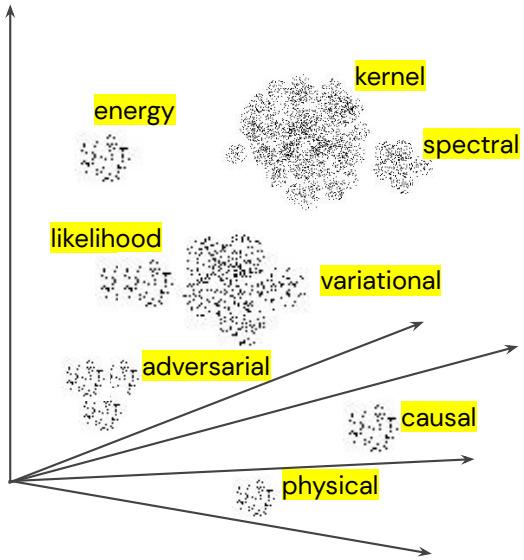
General representation learning without labels is **still largely unsolved**.

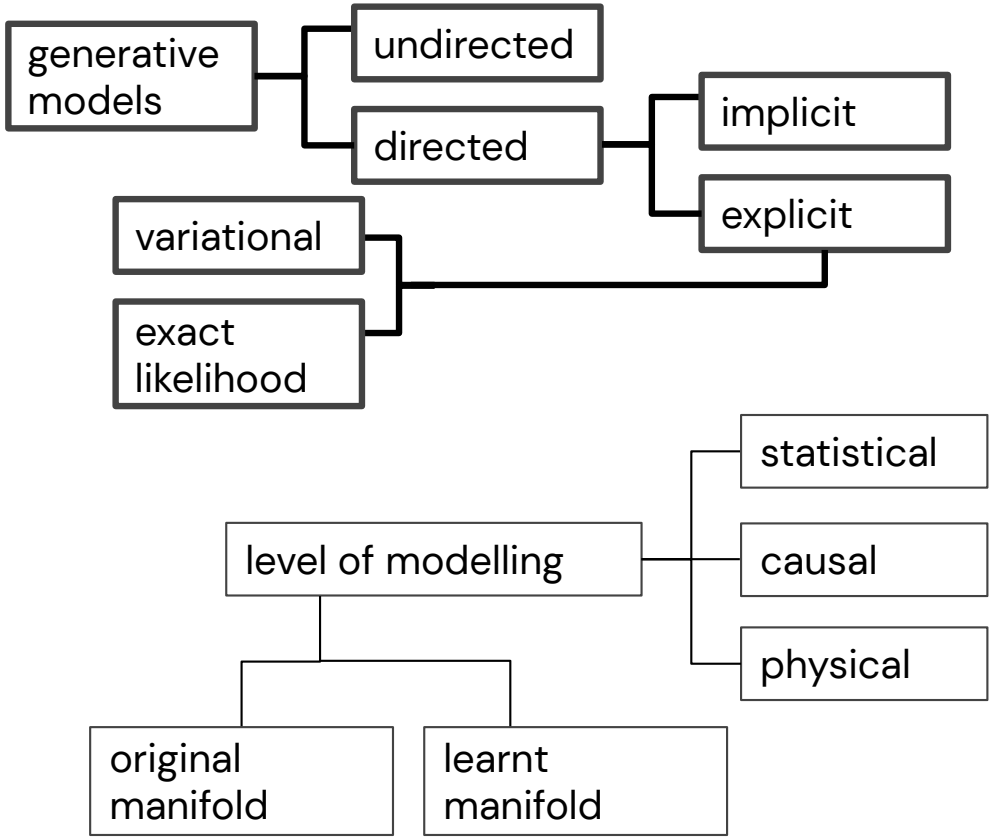
Recent advances, however, hold promise in finding increasingly general and useful representations.

DeepMind

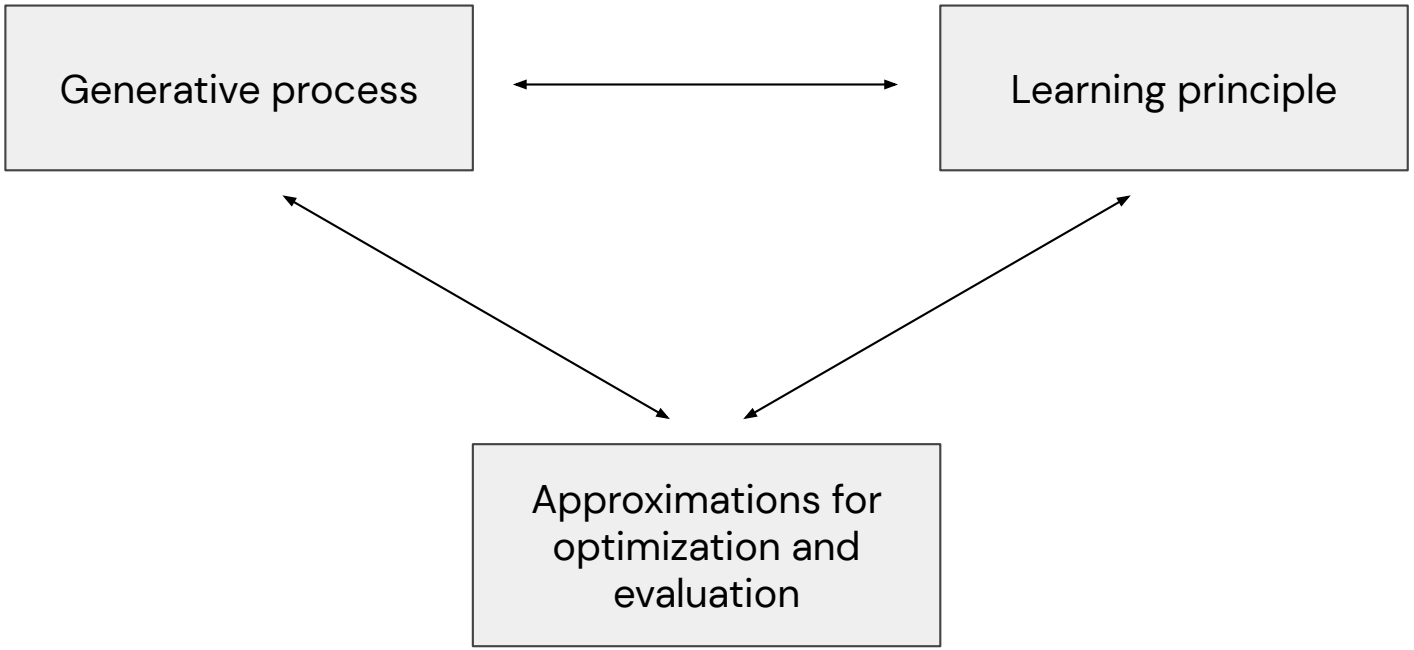
3 Landscape

Mapping out the landscape

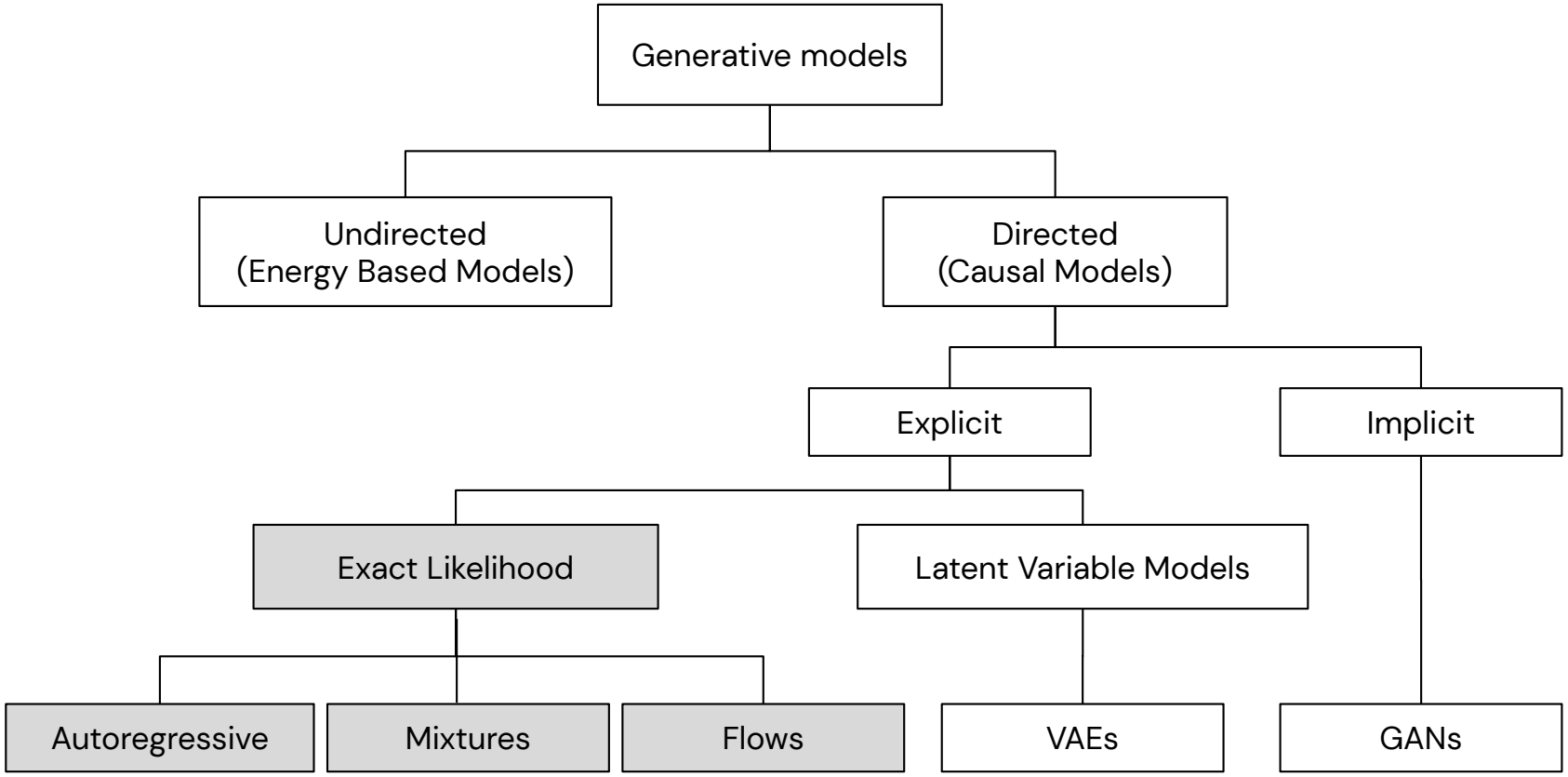




Foundations of generative models

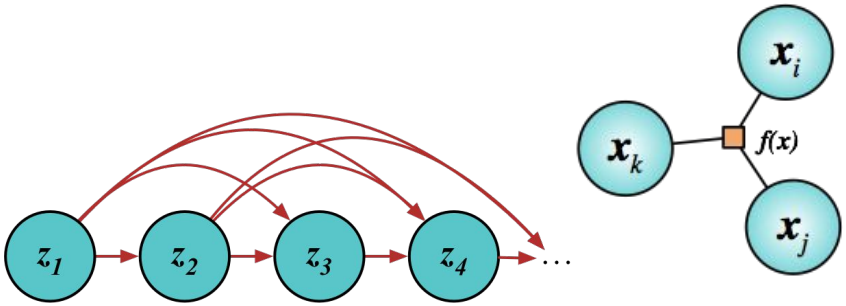


Mapping out the landscape of Generative Models



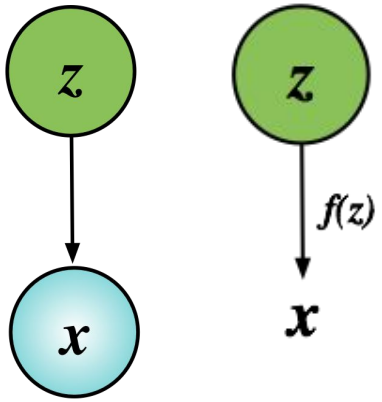
Types of Generative Models

Fully-observed models
Model observed data directly without introducing any new unobserved local variables.

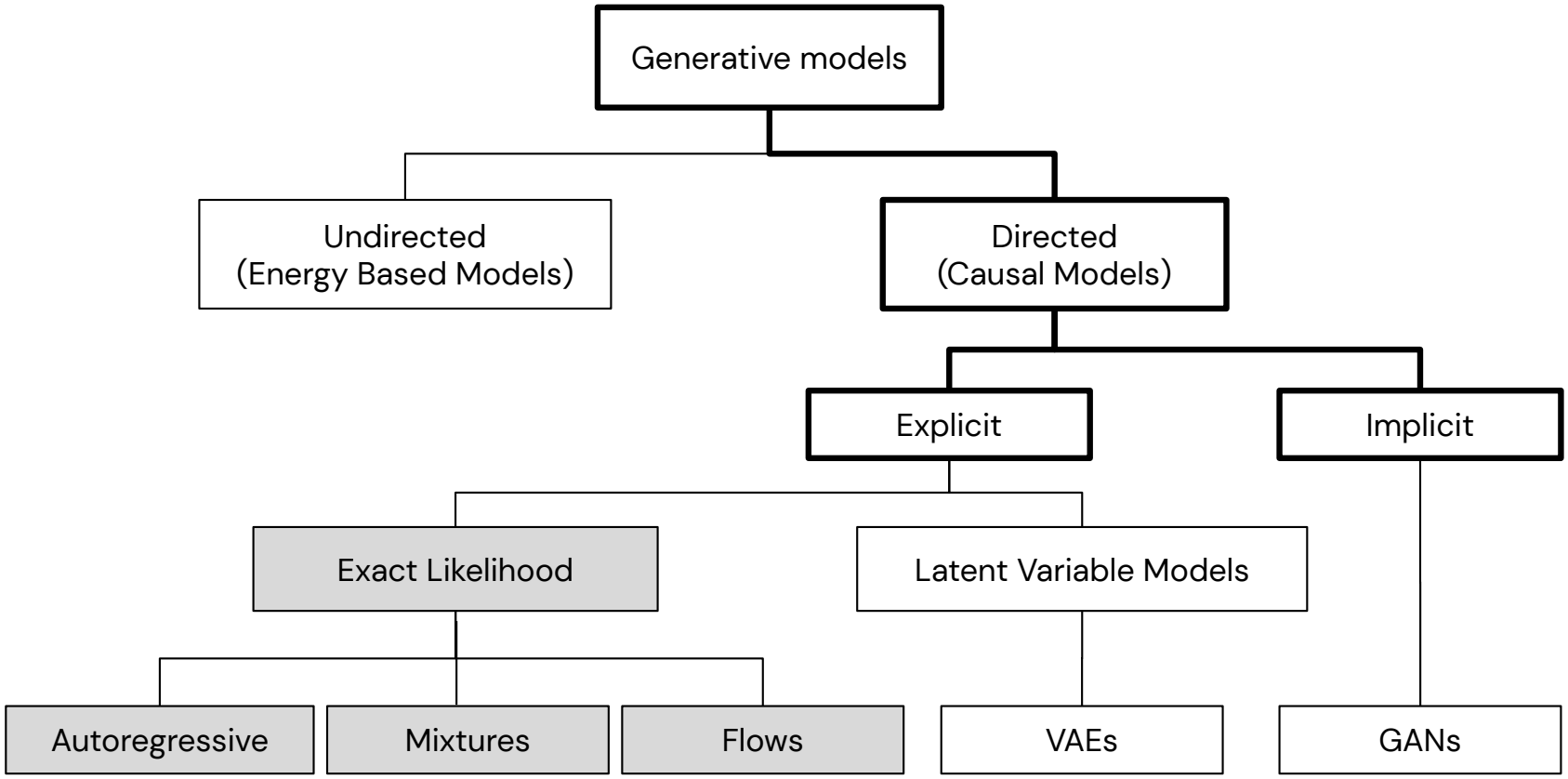


Latent Variable Models
Introduce an unobserved random variable for every observed data point to explain hidden causes.

- Prescribed models: Use observer likelihoods and assume observation noise.
- Implicit models: Likelihood-free models.



Mapping out the landscape of Generative Models



Foundations: density estimation & divergences

- Given an empirical density **p(x)** represented by a collection of iid samples **$\{x_1, ..., x_N\}$**
- Our goal is to modify the parameters **θ** of a parametric density **$q_\theta(x)$** so that it gets "closer" to **p(x)**
- But what means "closer"?

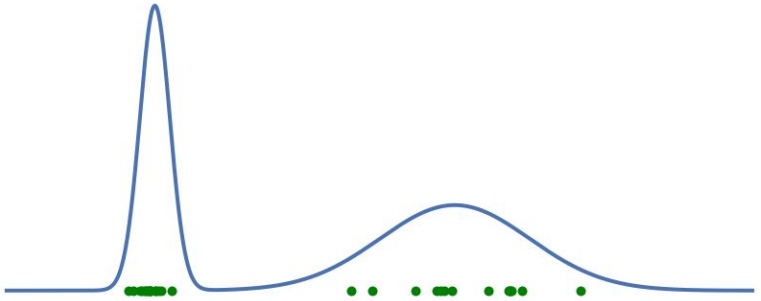


Image credit: Mihaela Rosca

Let $\mathcal{H}$ be the space of probability densities of interest (e.g. $\mathcal{H} = \{p : \mathbb{R} \to \mathbb{R}^+, ||p|| = 1\}$).
A *probability divergence* $\mathrm{D}(P;Q)$ is a map $\mathrm{D}: \mathcal{H}^2 \to \mathbb{R}$ such that:
For any P and Q, $\mathrm{D}(P;Q) \geq 0$
For any P and Q, $\mathrm{D}(P;Q) = 0 \Leftrightarrow P = Q$

Foundations: density estimation & divergences

Want to learn more?

Invariance, Rezende, Posts on ML, Math and Physics 2018
https://danilorezende.com/2018/07/12/short-notes-on-divergence-measures/

Each divergence will emphasize different aspects of the learned density

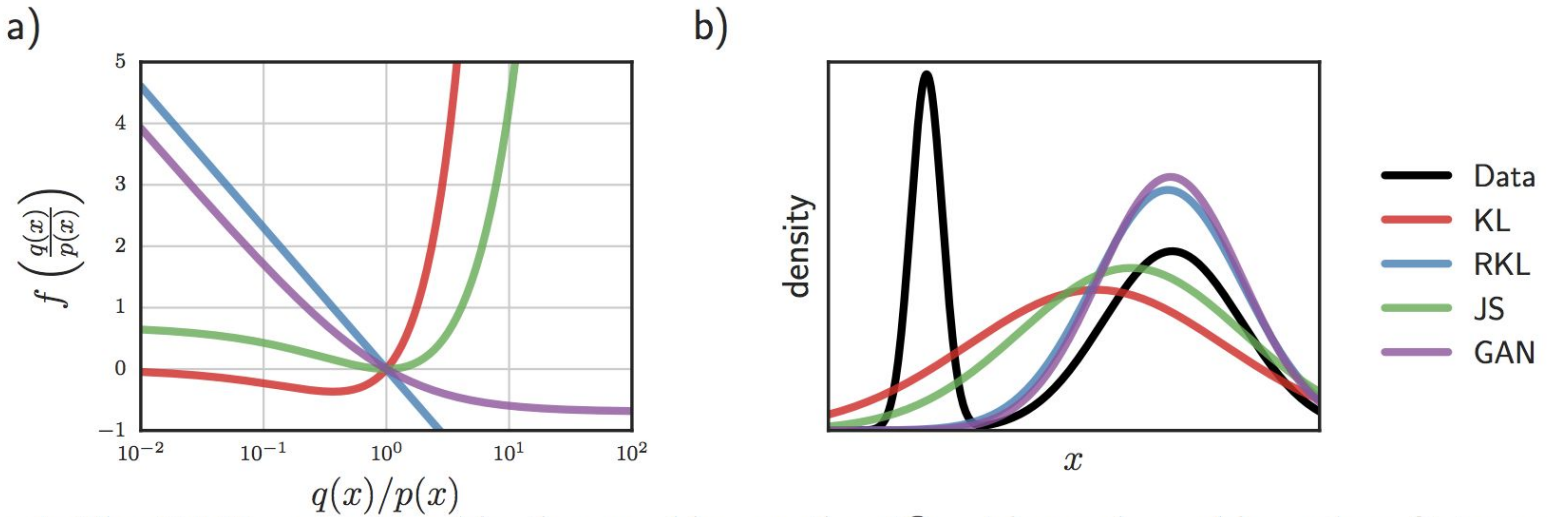
Name	**Formula**	**Condition**
f-divergences	$D(p;q) = \mathbb{E}_q[f(\frac{q}{p})]$	strictly convex f
Relative entropy (KL)	$D(p;q) = \mathbb{E}_q[\ln \frac{p}{q}]$	
Jensen-Shannon (JS)	$D(p;q) = \frac{1}{2}\mathrm{KL}(p;m) + \frac{1}{2}\mathrm{KL}(q;m)$	$m = \frac{1}{2}(p+q)$
Stein divergence	$D(p;q) = \sup_f \mathbb{E}_q[\nabla \ln pf + \nabla f]^2$	where $\int \nabla(pf) = 0$
Energy distance	$D(p;q) = \mathbb{E}[2\|x-y\| - \|x-x'\| - \|y-y'\|]$	$x, x' \sim p; y, y' \sim q$
Wasserstein distance	$D_\alpha(p;q) = [\inf_\rho \mathbb{E}_\rho[\|x-x'\|^\alpha]]^{\frac{1}{\alpha}}$	$\int dx\rho(x,x') = q(x')$ and $\int dx'\rho(x,x')$
Max-min dis. (MMD)	$D(p;q) = \sup_f(\mathbb{E}[f]_p - \mathbb{E}[f]_q)$	f continuous and bounded

Foundations: density estimation & divergences

Want to learn more?

Improved generator objectives for GANs, Poole et al, NeurIPS Workshop on Adversarial Training 2016

Each divergence will emphasize different aspects of the learned density



Foundations: density estimation & divergences

Want to learn more?

Estimating divergence functionals and the likelihood ratio by convex risk minimization, Nguyen, et al, IEEE Transactions on Information Theory 2010

A general bound on f-Divergences

Requires knowledge of **p(x)** and **q(x)**

$$D_f[p;q] = \mathbb{E}_q\left[f\left(\frac{p}{q}\right)\right] = \sup_{\phi \in L^2} \mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]$$

Only requires samples from **p** and **q** and evaluation of **ϕ(x)**

This observation is the key insight of GANs, allowing us to train models from which we can sample from but not evaluate its likelihood

Integral Probability Metrics

$$\mathcal{M}_f(p, q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{p(x)}[f] - \mathbb{E}_{q_\theta(x)}[f] \right|$$

f sometimes referred to as a **test function, witness function or a critic.**

Many choices of ***f*** available: classifiers or functions in specified spaces.

$\|f\|_L < 1$
Wasserstein

$\|f\|_\infty < 1$
Total Variation

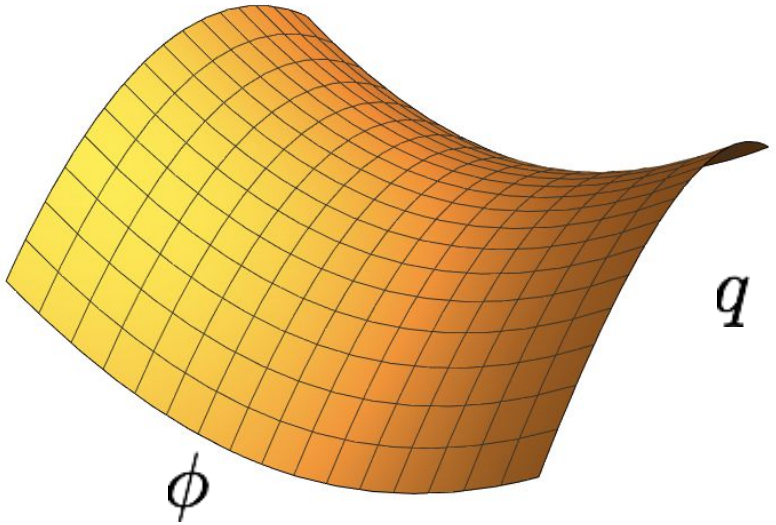
$\|f\|_{\mathcal{H}} < 1$
Max Mean Discrepancy

$\left\| \frac{df}{dx} \right\|_L < 1$
Cramer

Foundations: density estimation & divergences

A general bound on f-Divergences

$$\min_q D_f[p;q] = \min_q \max_\phi \mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]$$

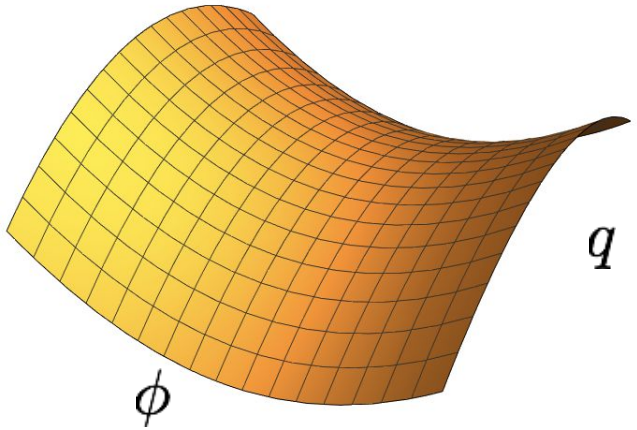


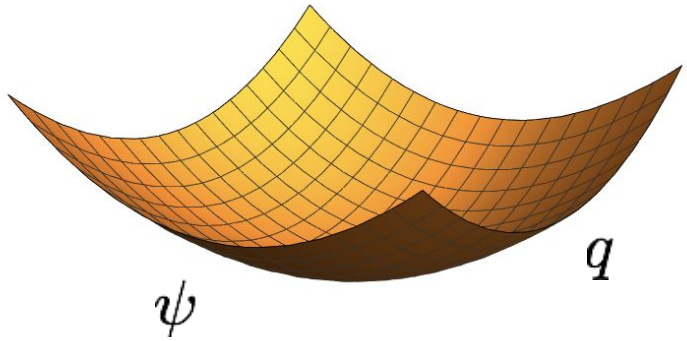
Foundations: density estimation & divergences

Upper and lower bounds on f-Divergences

$$f(u) = u \ln u$$

$$\boxed{\mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]} \leq D_f[p;q] \leq \boxed{-\mathrm{ELBO}_\psi[p;q]} + cst$$



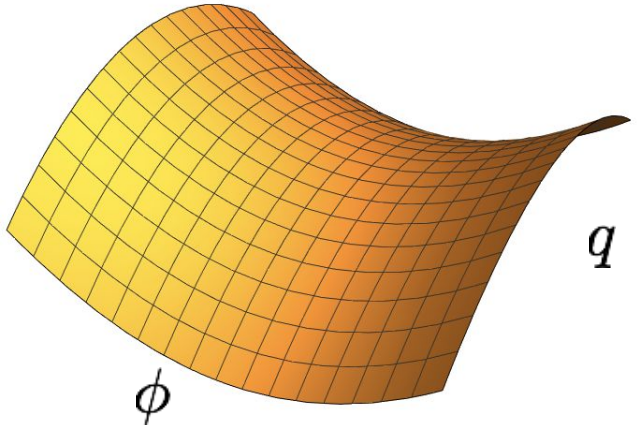


Foundations: density estimation & divergences

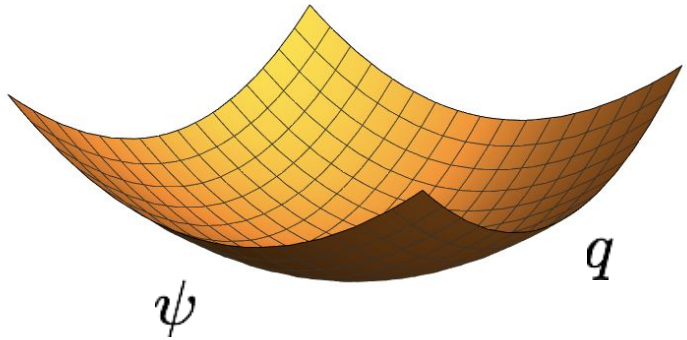
Upper and lower bounds on f-Divergences

$$f(u) = u \ln u$$

$$\boxed{\mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]} \leq D_f[p;q] \leq \boxed{-\mathrm{ELBO}_\psi[p;q]} + cst$$



GANs



VAEs

Foundations: density estimation & divergences

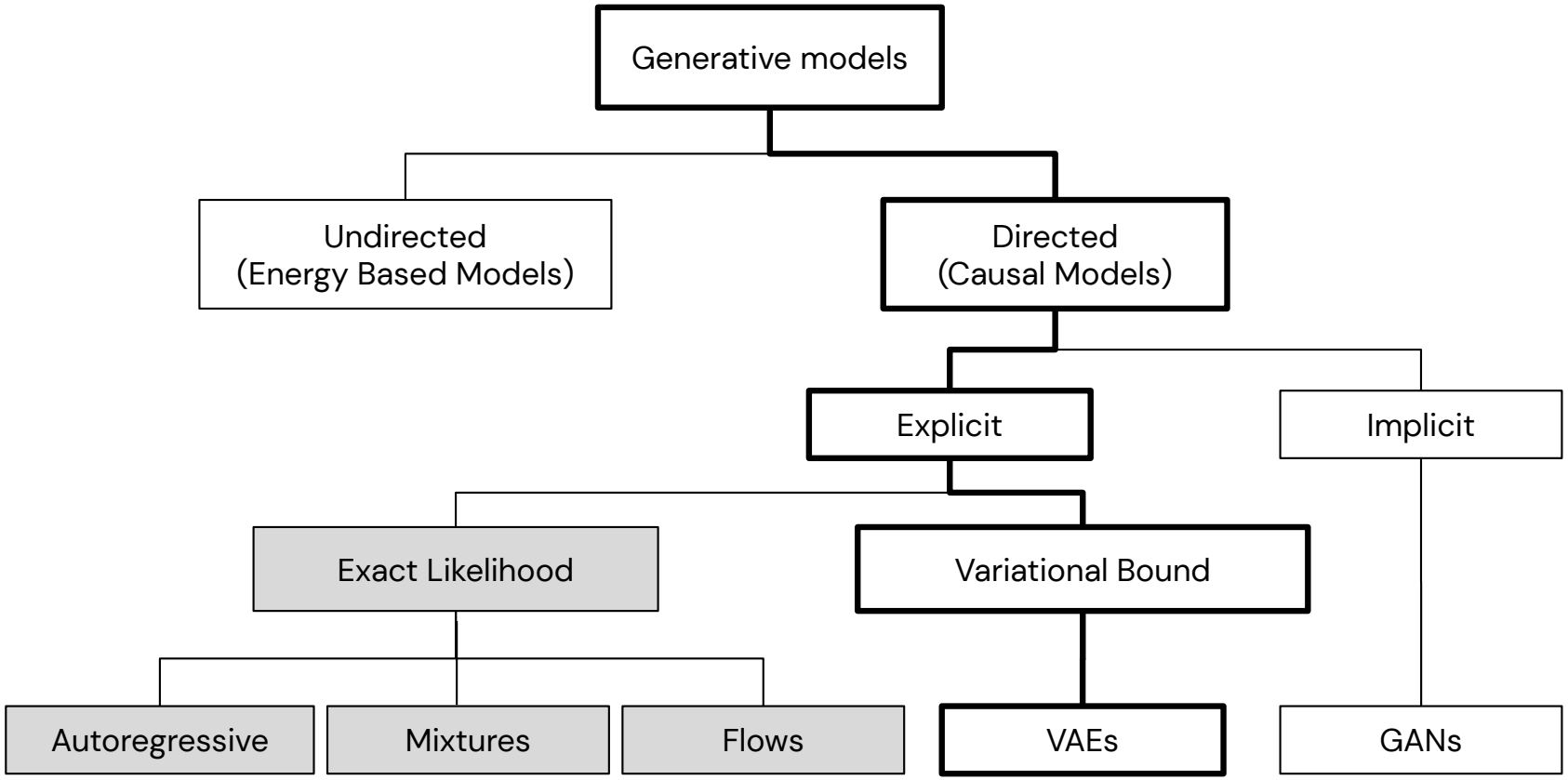
Want to learn more?

Improved generator objectives for GANs, Poole et al, NeurIPS Workshop on Adversarial Training 2016

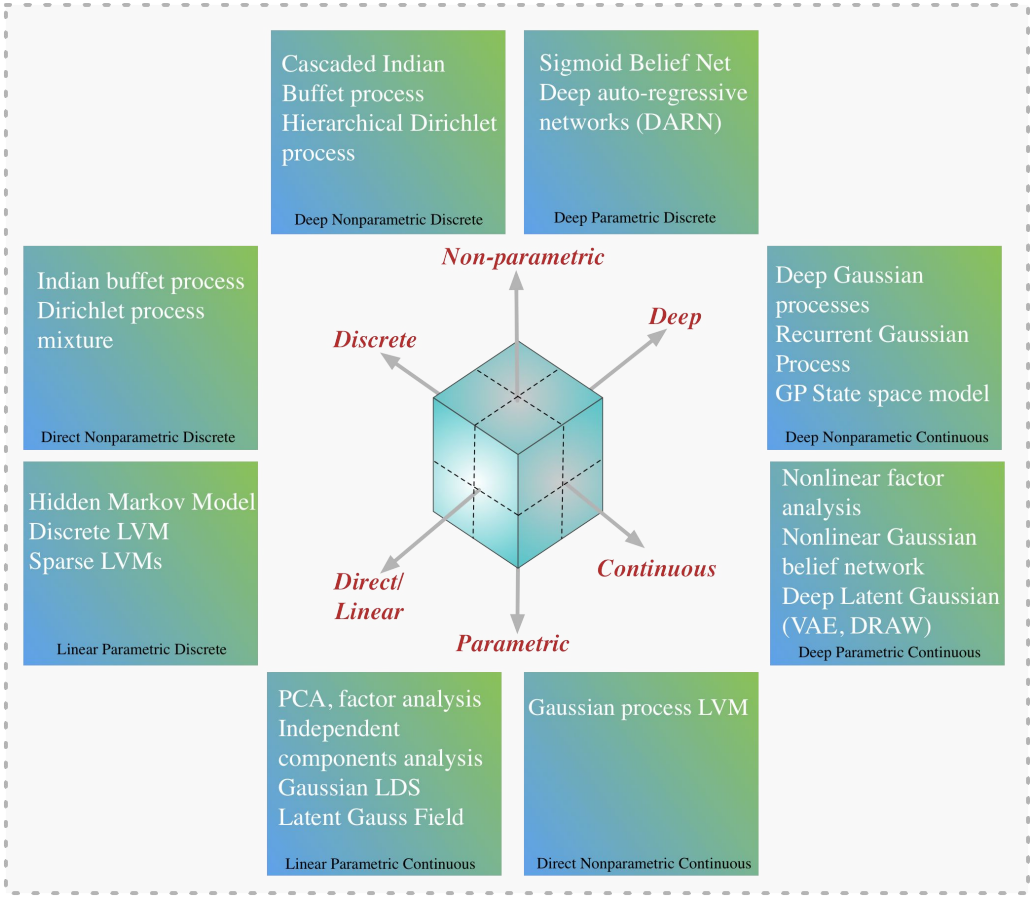
Many GAN models are particular cases of f-Divergences optimization

Name	Generator f-divergence (f_G)	Generator objective (minimized)
GAN-standard	$\log(1+\frac{1}{u})$	$\log\left(1+e^{-V(x)}\right) = -T(x)$
GAN-RKL	$-\log u$	$-V(x)$
GAN-KL	$u \log u$	$V(x)e^{V(x)}$
GAN-α	$\frac{1}{\alpha(\alpha-1)}\left(u^\alpha - 1 - \alpha(u-1)\right)$	$\frac{1}{\alpha(\alpha-1)}\left(e^{\alpha V(x)} - 1 - \alpha(e^{V(x)}-1)\right)$

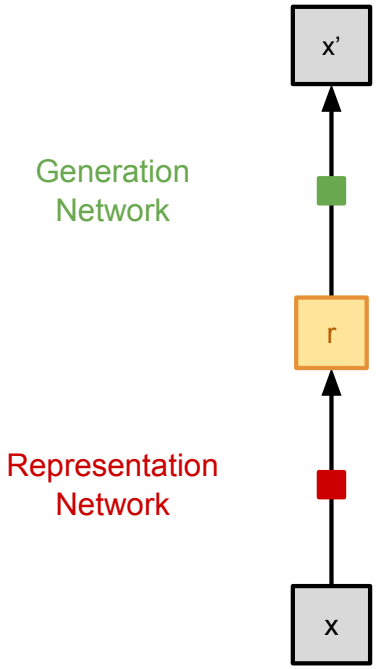
Mapping out the landscape of Generative Models

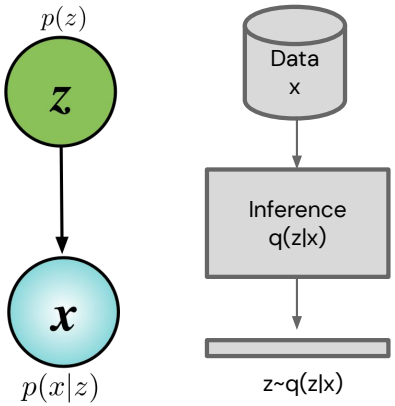


Spectrum of Latent Variable Models

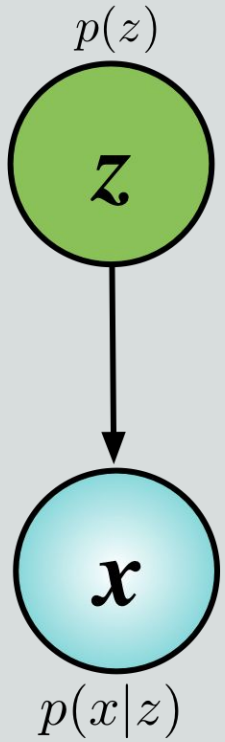


Representing Models: Computational graphs vs plate notation





Latent Variable Models



$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$
$$\mathcal{D} = \{x_i\} \quad i \in \{1, ..., N\}$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz = \log \mathbb{E}_{p(z)}[p_\theta(x|z)]$$

$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$